Harmful Algal Blooms

Timothy I. McLean[1], Geoffrey A. Sinclair[2]
[1]Department of Biological Sciences, University of Southern Mississippi, Hattiesburg, MS, USA
[2]DeFelice Marine Center, Louisiana Universities Marine Consortium, Chauvin, LA, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Alexandrium
Karenia brevis
Pseudo-nitzschia
Cyanobacteria
Future Directions
Bibliography

Glossary

Bioaccumulation The retention of substances, various phycotoxins in this context, in the living tissues of organisms at concentrations that are higher than are found in the environment.

Cyanobacteria A phylum of bacteria, once called blue-green algae, that appears in the fossil record as far back as 3.5 billion years all species are photosynthetic, but different species have adapted to aquatic (fresh or marine) or terrestrial (soil, bare rock, or symbiotic associations with plants) environments some species can form colonies, and some colony-forming cells can specialize or differentiate, e.g., to form nitrogen-fixing heterocysts.

Depuration The cleansing or purification of impurities (toxins) from the tissues of the body.

Diatoms A large group of microalgae distinguished by having siliceous cell walls called frustules single cells range in size from 2 μm to over 2 mm and are unicellular, often forming colonies of many cells they contain the pigments chlorophylls a and c, betacarotene, fucoxanthin, diatoxanthin, and diadinoxanthin they constitute the predominant fraction of the phytoplankton community in fresh water and marine environments, found in the water column and in the benthos, being both epiphytic and edaphic.

Dinoflagellates A large group of motile microalgae differentiated from other algal groups by their thecal morphology (cell covering) having either an armored form or a smooth form armored forms contain distinctly ornamented polysaccharide plates they contain two dissimilar flagellae, and species may be exclusively heterotrophic, mixotrophic, or exclusively phototrophic – those that are capable of photosynthesis contain the pigments chlorophylls a and c, dinoxanthin, diadinoxanthin, and either fucoxanthin or peridinin as their major carotenoid, depending upon the species.

Eutrophication The introduction of excessive levels of nutrients such as nitrates and phosphorous into an aquatic ecosystem.

Harmful algae A small fraction of algal species that can have disruptive or devastating effects to an ecosystem, public health, and/or various economies the harmful nature of different species of algae can be manifested in a variety of ways but primarily through the production of one or more phycotoxins.

Phycotoxins Toxic molecules produced by harmful algae collectively, these molecules are chemically diverse and can produce an array of effects in intoxicated organisms ranging from gastrointestinal, neurotoxic, hepatotoxic, carcinogenic, and more.

Trophic transfer The transfer of energy and/or substances, e.g., toxins, from one (trophic) level of a food chain to another, typically higher, (trophic) level within the food chain.

Robert A. Meyers (ed.), *Encyclopedia of Sustainability Science and Technology*, DOI 10.1007/978-1-4419-0851-3,
© Springer Science+Business Media, LLC 2012

Definition of the Subject and Its Importance

Harmful algal blooms (HABs) pose threats to the environment, public health, and a variety of commercial interests and industries. A single bloom can lead to devastating outcomes, including large mortalities of marine organisms (e.g., fish kills); toxic contamination of filter-feeding organisms such as bivalve shellfish that subsequently enter the market for distribution to consumers; economic hardships for fisheries, aquaculture, and recreational- and tourism-related industries; and a compromised quality of life for people living or working along affected shorelines. Depending upon the size of the bloom, its duration, and the number and types of impacts produced, a single bloom can generate multimillion-dollar losses spanning from local to international economies (see ► Oceans and Human Health, Social and Economic Impacts for additional information on HAB-related economic impacts).

As well as a current concern, the issue of HABs will continue to grow in the future, as HABs are being reported at higher frequencies, lasting for longer periods of time, covering larger areas of shoreline or acreage of water, and spreading to regions that historically have not experienced blooms [1–8]. A number of factors have been cited as contributing to this global increase, but each algal species responds to a different set of variables. For many species, degraded water quality is one of the major reasons for the increase in HABs and HAB effects [9, 10]. As eutrophication, urbanization, and climate change continue to escalate, the increasing impact of HABs is a threat to the economic sustainability of many coastal and marine industries, public health, and the health of natural resources.

Introduction

“Algae” is not a true taxonomic designation. Algae are polyphyletic, i.e., species under this umbrella term can be found on numerous branches within the eukaryotic lineage. The diversity of organisms within this assemblage make it difficult to provide a concise definition of algae (reviewed in Ref. [11]), but in general terms algae can be described as simple (i.e., lacking the structures and vasculature characteristic of land plants), mostly aquatic, photosynthetic, and ranging in form from unicellular (microalgae) to multicellular (macroalgae). For historical and physiological reasons, this artificial grouping also includes the prokaryotic cyanobacteria, formerly known as blue-green algae. The inclusion of cyanobacteria is further strengthened in the context of discussions about harmful algae (see below).

Algal blooms denote an increase in the population of a single species of algae within a given area. The increase may be a result of an increased growth rate or the accumulation/aggregation of cells due to, respectively, an infusion of nutrients or physical features such as fronts, a stratified water column, or current entrainment [9, 12, 13]. Algal blooms can form for a multitude of reasons, and the reasons for and the characteristics of a bloom vary by the species. Algal blooms occur worldwide in fresh, brackish, and marine environments. If the cell concentration becomes sufficiently dense, the bloom may discolor the water to produce a “red tide,” although, depending on the organism, the water may actually appear green, brown, or gold, instead of red. Although the term red tide has taken on a negative connotation, the majority of red tides occur as natural, seasonal phenomenon that produce beneficial rather than harmful consequences. For example, in the Southern Ocean, successions of blooms produced by different algal species during the spring and early summer months anchor food chains that feed a multitude of organisms from birds to whales [14]and sustains an important fishery that nets >100,000 t of krill per year [15]. Less than a tenth of the 3,400–4,100 identified algal species are known bloom formers, and less than a quarter of those (2% of the total) have the capacity to form a harmful algal bloom [16, 17].

Algal blooms are designated as harmful when they have detrimental effects to other organisms or ecosystem functions either directly or indirectly. In some cases, the mere presence of a high number of cells in the water column increases cell–cell collisions that may damage or kill other planktonic organisms. For example, some microalgae, particularly diatoms with extensive silicate setae, can cause physical trauma when they pierce or tear the cellular structure of other microbes or the gills of fish and crustaceans [1, 18, 19]. Indirectly, harmful effects include changes in resource competition or decreasing ecosystem health by producing biomass whose decomposition fuels the creation of hypoxic or anoxic conditions [20, 21]. Other harmful species may produce allelopathic molecules [22], allelochemicals [23], high amounts of ammonium [24], or extracellular polymeric secretions that create viscous or mucilaginous barriers in the water [25, 26]. The majority of our attention, however, has been focused on those algae that produce one or more toxic molecules referred to as phycotoxins. Direct contact with, inhalation of, or ingestion of the phycotoxins by humans (or other animals) can cause a range of effects from mild irritation to severe gastrointestinal distress to death. Some of the more hazardous phycotoxins can have hemolytic, hemoagglutinating, neurotoxic, or hepatotoxic effects, or they can promote the formation of tumors, e.g., in the colon (see below for specific examples and references). As described below, these toxins can vector through food webs as they pass from one trophic level to another. During these trophic transfers, toxins may bioaccumulate in tissues or undergo changes in potency. The effects of phycotoxins in the natural environment can lead to massive mortalities of fish, birds, marine mammals, and other less well-characterized marine organisms creating miles of coastline and hectares of water littered with dead animals.

A growing concern about HABs is the real or perceived threat that they are increasing year to year as evidenced by: (1) areas that have historically experienced blooms sporadically, or even rarely, are seeing an increase in frequency; (2) when blooms do occur, they are lasting longer and/or covering larger expanses of coastline; (3) new species of toxic algae are being discovered; (4) species that were previously described as nontoxic are now showing the capacity to be toxic; and (5) the biogeography of certain species is expanding [1–8]. Certainly, our ability and impetus to detect harmful species and their respective toxins has increased over the last 15 years, but the increased awareness and concern about HABs cannot fully explain the explosion in the number and extent of blooms around the world. The most likely culprits, although not for each species, are the introduction of species to new environments through mechanisms such as transport in ballast water and the alteration of the natural environment through mechanisms such as urbanization of our coastlines, diversion of natural water flows, draining of wetlands, introduction of new or increased sources of nutrients from aquaculture practices and runoff of urban and agricultural regions, and climate change. Any of these factors, or combinations of these factors, may produce an environment conducive for HAB species to bloom.

The following sections will highlight four of the better-studied HAB taxa (two genera: *Alexandrium* and *Pseudo-nitzschia*, one species: *Karenia brevis*, and one phylum: cyanobacteria) in terms of why they are harmful, the economic costs and types of impacts associated with each respective bloom, the methods used to detect the cells and/or associated toxin(s), and what is known in regard to what initiates and/or maintains the blooms. For comprehensive overviews of other HAB species or more information about the taxa described herein, refer to recommended readings at the end or the embedded reviews in each section.

Alexandrium

Description

The dinoflagellate genus *Alexandrium* currently contains greater than 30 species [27]. Members of the genus are armored, have a defined cellular and nuclear morphology, occur as either single cells or as chains of two or more cells, and are widely distributed in temperate regions of the Mediterranean and northern and southern Atlantic and Pacific Oceans [27]. Molecular data support the monophyletic grouping of *Alexandrium* spp. [28]. The same data challenge the division of species within the genus, however, as DNA-based assays show that cells tend to cluster based on geography rather than morphospecies designations [29–33]. Twelve *Alexandrium* spp. are confirmed to produce paralytic shellfish toxins ([34] and references therein). Members of the *Alexandrium tamarense* morphospecies complex (*A. tamarense*, *A. catenella*, and *A. fundyense*) are responsible for the majority of blooms along northeastern and northwestern North American coastlines.

Toxicity

Paralytic shellfish toxins (PSTs) consist of saxitoxins and at least 24 saxitoxin-like congeners (reviewed in Ref. [35]). (It is important to note that, while not discussed here, other dinoflagellates, principally *Gymnodinium catenatum* and *Pyrodinium bahamense* var. *compressum*, and some cyanobacterial species can also produce PSTs (reviewed in Ref. [35]).)

The potency of different PSTs ranges over a scale of one to three orders of magnitude [36], and those with a sulfocarbamoyl group are the most toxic (reviewed in Ref. [35, 37]). The name of this grouping of toxins derives from the syndrome, paralytic shellfish poisoning (PSP), which is produced as a result of consuming seafood, primarily bivalves, containing PSTs. PSTs, such as saxitoxin, bind voltage-gated sodium channels on the membranes of electrically excitable cells such as neurons and block the flow of sodium ions through the channel [38, 39]. The blockage prevents depolarization of the cell [40], which causes a range of gastrointestinal and neurological symptoms that vary with the exact composition of various PSTs and the susceptibility of the individual. Symptoms include tingling around the mouth, numbness of the extremities, weakness, nausea, and vomiting, among others. If a sufficient dose is ingested, neurons will be unable to stimulate muscle contractions. Complete muscle relaxation leads to paralysis, and if the paralysis extends to the muscles controlling breathing, individuals can die from asphyxiation [41].

PSTs can vector through marine food webs, having been found in zooplankton and zooplanktivorous fish (reviewed in Ref. [42]). The bioaccumulation of PSTs in these lower trophic level organisms has been implicated in mortalities of large numbers of fish, oceanic and coastal birds, and marine mammals, including humpback and right whales, which consume whole fish [4, 43–48]. Human intoxication from eating fish is rare since the PSTs accumulate in the viscera, which is typically not eaten by people. The most common route for human intoxication is through consumption of contaminated bivalves (e.g., mussels and clams), although, in rare cases, other organisms (e.g., gastropods, crustaceans, and fish) have served as vectors [49, 50]. Shellfish are indiscriminate filter feeders; they can consume large quantities of microalgae, are relatively immune to the PSTs, and bioaccumulate the toxins in their tissues. Different bivalve species have different tolerances to PSTs, and having a high tolerance allows an organism to accumulate more toxins. An example of an organism with a high tolerance is the blue mussel (*Mytilus edulis*). In Alaskan waters, *M. edulis* can bioaccumulate greater than 20 mg of saxitoxin/100 g mussel tissue – more than enough saxitoxin to kill a person [51]. Typically, 1–2 months after moderate blooms have terminated, the majority of bivalves will have depurated the toxins and will be considered safe for harvest and consumption. By contrast, butter clams bind to saxitoxin with high affinity and can remain toxic for up to 3 years after a bloom has occurred [51]. In some bivalves, PSTs can undergo biotransformation, possibly becoming even more toxic than the original source toxins [52].

Economics

PSTs are heat and acid stable, so cooking not only does not detoxify seafood but also may convert some congeners to more toxic forms. The only way to avoid PSP incidents is to prevent the harvesting of contaminated marine organisms or process them in such a way to reduce or minimize toxicity. For example, adductor muscles of sea scallops, which typically contains very low concentrations of PSTs, are considered safe if separated from the rest of the scallop while at sea before being frozen [53]. If the whole organism is frozen, then PSTs may leech from other tissues into the adductor muscle. For a review of other methods to reduce toxicity in seafood, see [53].

Alternatively, the only mechanism for determining whether to harvest/not harvest or to allow shellfish beds to be open or closed is to monitor for the presence of PSTs in the relevant shellfish tissues. Unlike other HAB species, detection and quantifying of PSP-producing microalgal cells are not sufficient for such management decisions – the collection of tissues and analysis for PSTs is required. These steps add costs to monitoring efforts performed by the relevant agencies. Costs for monitoring can become expensive especially for states with extensive coastlines. Alaska, e.g., only samples random sites and, therefore, discourages recreational harvesting, which leads to a severe underutilization of the endemic shellfish resources [8]. Of course, in a grand economic sense, the lost income and sales are, in part, balanced by reduced monitoring and public health costs.

In terms of economic consequences of PST-producing blooms, only a few case studies have been performed to quantify in dollars how extensive these impacts are. In the Pacific northwest, an *Alexandrium* bloom in 1997 created toxic oysters along the coast of Washington state that resulted in approximately

$9.45 million of lost revenue during a 2- to 3-week industry closure. An extended closure (2 months) of shellfish beds in South Puget Sound resulted in about $27 million in losses (Nosho in Ref. [54]). (Note: All dollar figures throughout this article are represented in 2011 values.) A similar bloom that closed beds along Washington, Oregon, and California state coastlines for a month in 1980 cost oyster growers $2.5 million [55]. During the same year in the northeastern United States, a bloom of *A. tamarense* closed the entire coastline of Maine to shellfish harvesting. The value of the lost harvest was placed at $6.6 million, and the total economic impact was placed at $20 million [56]. In 2005, a massive bloom extended over the majority of the New England coastline, resulting in $3.1 million/week in lost revenue for the affected seafood fisheries [57]. Later analysis determined that the bloom impacted the state of Massachusetts by greater than $20.7 million and at least $2.7 million to the state of Maine [58].

Detection

A number of molecular techniques have been developed to rapidly and correctly identify cells as a particular *Alexandrium* species. The majority of the efforts to differentiate species have focused on molecular sequence data, such as the sequence of the ribosomal RNA gene for either the small or large ribosomal subunit or the internal transcribed spacer [29, 30, 32, 59–61]. Depending upon the marker or technique used, the number of defining ribotypes for *Alexandrium* varies from 5 to 11 [30, 32, 60]. Despite the discrepancy in the number of ribotypes, the results of these assays show that *Alexandrium* cells from the same geographic region are more closely related to each other than to cells from different regions, even if the cells from the different areas are of the same species. The conclusion drawn from the data is that *Alexandrium* is monophyletic, but geographically isolated populations evolved independently subsequent to global dispersion [31].

More effort has gone to developing methods for detecting and quantifying PSTs. The standard assay is the mouse bioassay in which acid extracts of tissue samples from intoxicated organisms are injected into three mice and the average time to death is recorded [62, 63]. The results are converted into a concentration of saxitoxin, or saxitoxin equivalents, which then serves as the basis for determining whether a particular shellfish site stays open or is closed. The current action level in the United States set by the Food and Drug Administration (FDA) is 0.8 ppm (80 μg saxitoxin/100 g tissue) [64]. The action level in other countries may vary slightly and can be found in FAO Marine Biotoxins Paper 80 [65].

Despite monitoring and informing the public of the dangers of PSP, there are still sporadic poisoning incidents from recreational or subsistence harvesting in closed or unmonitored harvest sites [35]. The use of the mouse bioassay has always been fraught with concerns and problems (e.g., poor specificity for different congeners, low sensitivity, high variability, and the sacrifice of a large number of animals), and a number of alternative techniques have been and continue to be developed. Some of these techniques include neuroblastoma tissue culture bioassays [66, 67], a hemolysis assay [68], receptor-binding assays [69–71], enzyme-linked immunosorbent assays (ELISAs) [72, 73], and electrochemical biosensors [74–77]. Most of these techniques have offered improvement in some areas, but they have not been sufficiently developed or tested to replace the mouse bioassay. Of those listed, the ELISAs have received the most attention in terms of lowering the limit of detection and increasing broad-spectrum detection or developing congener-specific antibodies, e.g., [78–82]. An offshoot of the ELISA technique has the development of lateral flow immunochromatography assays [83, 84]. One such assay, the quick (less than 20 min), one-step MIST Alert™ (Jellet Rapid Testing Ltd.), uses a combination of highly specific antibodies to overcome cross-reactivity issues that have plagued immunoassay development. MIST Alert™ has been field tested, and the biggest concern with its use is the high rate of false positives [35, 85, 86]. The US FDA has approved the test kits for use in the field to prescreen samples – negative samples do not need further analysis whereas positives must be confirmed via mouse bioassays [35]. The prescreening will save time, money, and animal lives. A lot of effort has also been expended in regard to developing liquid chromatography-based techniques. Originally developed by the US FDA [87], a number of derivations and modifications have been implemented including tandem linkage to mass spectrometry and pre- and post-column oxidation of the toxins [75, 76, 88–95]. A particularly useful development was the creation and commercialization of PST standards for use in analyses such as liquid chromatography [96, 97]. Recently, the use of liquid chromatography with post-column oxidation and fluorescence detection has been approved as an official method of analysis by the AOAC [98]. The establishment and validation of this methodology is often a prerequisite for adoption by regulatory agencies and may pave the way for the replacement of the mouse bioassay as the standard assay for detection and quantitation of PSTs.

Origins and Nutrient Interactions

PSP events were reported in Canada over 100 years ago (Ganog in Ref. [99]), but the first documented case in the United States was not until 1958 [100]. In the New England area of the United States, specifically the western Gulf of Maine, PSP events were relatively rare, but they are increasing in frequency and becoming more widespread [101–103]. Similarly, *Alexandrium* and PSP events are now common in historically pristine Pacific waters such as those around Alaska and northern Japan [101]. From these observations, in conjunction with the results from laboratory and field experiments, the spreading of *Alexandrium* blooms and toxifying events can best be explained by examination of cyst deposition and cell transport via storms and currents. A combination of extensive spatial and temporal sampling of cyst seed beds, environmental monitoring and observations, and laboratory experimentation has produced a biophysical model of *Alexandrium* in the Gulf of Maine [57, 104]. The model successfully recreated ("hindcasted") the formation and movement of the 2005 *Alexandrium fundyense* bloom that started in the northern Gulf of Maine and spread southward as far as Long Island, New York [57, 104]. The model has highlighted the role of three factors of particular import for determining bloom impacts to the coastline. First, heavy rainfall and snow melt created substantial freshwater inputs. The inflow provided some nutrients to the coastal ecosystem, but it also helped stratify the water and provided buoyant plumes that helped transport cells in coastal currents. Second, upwelling-favorable winds shifted to become downwelling-favorable, which promoted transport of offshore blooms onshore. Lastly, bloom intensity is directly correlated to the abundance of cysts present in one or both of two seed beds – in the Bay of Fundy or offshore of Penobscot and Casco Bays [57, 104–106]. Cysts are deposited at the end of a bloom where they overwinter (i.e., lie dormant in the sediment) until the spring, at which time the warmer water temperatures and increased light, nutrients, and oxygen will cause the cysts to germinate [107]. Upon germination the resulting vegetative cells rise to the surface, and they can be pushed onshore under downwelling-favorable conditions and transported along the coastline by prevailing winds and currents [57]. As blooms spread to new areas, the seeding of those areas with cysts likely expands the range of blooms [55, 57].

Understanding the role and importance of these variables and the creation of a model provides significant value as a year to year predictor of bloom likelihood and bloom severity. The model has been applied to the bloom seasons since 2006 and has provided guidance for shellfish harvesters, stakeholders, and managers to make more-intelligent decisions about when and where to harvest shellfish, when and where to monitor for toxin, and how best to preserve public health [8]. In light of the importance of cyst seed beds, similar beds are being mapped and monitored in Puget Sound, Washington, as potential initiating sites [8]. The information also provides a warning against indiscriminant dredging activities that may unearth buried, dormant cysts. Almost all cysts (90%) are buried below the surface in anoxic conditions [108] – conditions that sustain their encystment unless resuspended. Dredging may "reawaken" these cysts to form a bloom or enrich a bloom with additional cells the following spring. The transporting and depositing of marine sediment that may contain cysts to new locations must be monitored lest these activities serve to seed new areas and spread the occurrence of *Alexandrium* blooms [102].

Karenia brevis

Description

The harmful dinoflagellate *Karenia brevis* is an athecate dinoflagellate, i.e., it does not possess any armored plates, and it is about 20 μm in diameter. The size of *Karenia brevis* varies over a diel cycle increasing in size as the day progresses [109]. *Karenia brevis* swims about 250 μm per second or about 1 m an hour. This swimming ability allows *Karenia brevis* to undergo a diel vertical migration, ascending during the day and descending at night [110]. The traditional explanation for this pattern is that cells are maximizing exposure to light during the day to support photosynthesis and maximizing exposure to nutrients at night. *K. brevis* does not, however, undergo a banded migration characterized by synchronous ascent and descent of the entire population, and therefore there appears to be other factors involved in its behavior other than external stimuli.

K. brevis has been intensively studied due to the socioeconomic impacts that blooms have on coastlines throughout the Gulf of Mexico but particularly on the west coast of Florida, where they now occur almost annually [111, 112]. From January to October 2006, Florida red tides killed thousands of fish and numerous manatees and caused incidents of human respiratory distress, which ultimately reduced local tourism. While cell densities of greater than 1×10^5 cells per liter near shore are responsible for these problems, the *K. brevis* populations appear to originate farther offshore in oligotrophic water columns [112, 113].

Toxicity

Karenia brevis produces a suite of cyclic polyether compounds referred to as brevetoxins (PbTxs). Brevetoxins are a type of neurotoxin that disrupts normal neurological processes by binding to voltage-gated sodium channels in nerve cells. Consumption of this toxin results in an illness described as neurotoxic shellfish poisoning [114]. For decades, brevetoxins produced by *K. brevis* have been implicated in animal mortalities throughout the food chain [115, 116]. Blooms not only result in massive fish kills, but brevetoxins also bioaccummulate in fish and result in mortality of bottlenose dolphins that regularly feed on those fish [116, 117]. Other mammals, like the protected manatee, experience mortality either by inhalation of aerosols or by consumption of contaminated sea grasses [115].

Unlike toxins produced by other species, brevetoxins can be aerosolized, affecting the human population along the coast [118, 119]. During a Florida red tide, aerosolized PbTxs produced by *Karenia brevis* cause both immediate and prolonged airway symptoms in humans, especially in those with preexisting airway disease (e.g., asthma). Environmental monitoring indicates that toxins remain airborne for up to four consecutive days. Repeated exposure to PbTx-3 can result in prolonged airway hyperresponsiveness and lung inflammation, and these pathophysiologic responses may contribute to the prolonged respiratory symptoms observed in humans after red tides [120].

Economics

Karenia brevis is a toxic dinoflagellate that is responsible for extensive ecological and socioeconomic impacts from Florida to Texas [121]. Brevetoxins, produced by *K. brevis*, have been implicated in mortalities of marine life including fish, dolphins, and manatees [122, 123], as well as causing respiratory illness in people and affecting local tourism [124–126]. The economic impacts in Florida have been estimated to range from $20 to $34 million per year [127] with losses incurred from restaurant closures to emergency room visits to shellfishery closures [128–131].

Detection

Despite the regularity of impacts of *K. brevis* blooms on the West Florida Shelf (WFS), predicting where and when *K. brevis* blooms occur remains problematic. The difficulty in predicting the occurrence of these populations resides in the phenomenon that dense populations of *K. brevis* (greater than 1×10^5 cells per liter) often appear more quickly than can be explained by growth rates alone [112, 132]. Increases in chlorophyll *a* biomass observed at frontal boundaries through satellite imagery can be tenfold greater over a 1- to 2-day period than can be explained by in situ growth [132]. Such unexplained increases have led to hypotheses of stealthy bloom initiation, defined as persistence of a population near the bottom of the water column [133] and below the first optical depth, therefore undetected by satellites over broad areas of the WFS [134]. Stumpf et al. [132] hypothesized that nutrients derived from the Mississippi River support similar near-bottom populations between the 25- and 50-m isobaths. As winds shift in the fall and promote upwelling conditions, these populations are subsequently advected onshore to areas of accumulation such as frontal zones [132].

Numerical models of *K. brevis* populations that are transported in the bottom Ekman layers replicate observations of in situ aggregation patterns in frontal zones [135, 136]. These models were based upon an extended Eulerian model of biologically and physically driven *K. brevis* population dynamics [137] and demonstrate the importance of incorporating the behavior and physiology of *K. brevis* into transport models. *K. brevis* undergoes a diel vertical migration (DVM) ascending just before sunrise and descending at night [138]. Recent observations of the diel vertical migration of *K. brevis* in mesocosms suggest that this diel vertical migration is to some degree under metabolic control [133, 139, 140]. Numerous scenarios of a biophysical model suggest that the trapping mechanism at frontal boundaries results from a combination of fluid advection and swimming behavior in response to internal metabolic state and ambient environmental stimuli (such as light and nutrients [135]).

Origins and Nutrient Interactions

Impacts are particularly troublesome in the region between Tampa Bay and Charlotte Harbor, a region that was labeled the red tide epicenter since *K. brevis* blooms were observed in 22 of the last 23 years compared to 5–10 outbreaks north and south of that region [127]. Blooms have been hypothesized to initiate 18–74 km offshore and in oligotrophic water columns [112, 113]. These oligotrophic water columns are characterized by nitrogen and phosphorus concentrations that are commonly about 0.1–0.2 μM within 2–4 km of shore in the Gulf of Mexico [141, 142] and less than 0.01 μM for NO_3^- and NH_4^+ [143]. Nitrate concentrations in both the surface and bottom waters at the 25-m and 50-m isobaths are often at or below 0.1 μM and are not able to sustain large populations of *K. brevis* associated with bloom events [113]. The importance of nutrients to the development and maintenance of bloom events has resulted in numerous hypotheses about potential sources of nutrients that may support populations of *K. brevis* [113, 144]. Potential nutrient sources for *K. brevis* in oligotrophic water columns include aerial deposition [113], N-fixation by *Trichodesmium* sp. blooms [145, 146], water column regeneration from grazers [147], near-bottom regeneration by diatoms responding to upwelling events [113], and regenerated nutrients derived from Mississippi River water [132]. Despite inputs of new nitrogen into the system, nitrate concentrations in both the surface and bottom waters at the 25-m and 50-m isobath are often at or below 0.1 μM [113]. Nutrients in the lower part of the water column, derived either from the sediment or from upwelling plumes below the pycnocline, may also be particularly important in supporting vertically migrating populations [148, 149].

Pseudo-nitzschia

Description

Diatoms are very proficient at utilizing available nutrients and, hence, at forming blooms [12]; however, relatively few diatoms are known to produce harmful blooms, and some species of the genus *Pseudo-nitzschia* are among them. *Pseudo-nitzschia* spp. are cosmopolitan, pennate diatoms found in nearly every major marine and estuarine environment on earth (polar to tropical, coastal to oceanic) [150, 151]. The growth of different *Pseudo-nitzschia* species varies based on each species' optimal ranges to a variety of environmental conditions. So, while *Pseudo-nitzschia* is detectable at the genus level year-round in many locales, the detectability is often a consequence of successions of multiple species in which any one species predominates based on the time of year or the prevailing conditions [152–154]. High levels of inter- and intraspecific variability have been detected at the genetic level [155–159]; this variability is often mirrored in regard to toxicity as well [160]. Twelve of the roughly 30 species of *Pseudo-nitzschia* are capable of producing the neurotoxin domoic acid [161–163]. Domoic acid-producing *Pseudo-nitzschia* species have produced significant blooms along the Pacific Northwestern coastlines of the United States and Canada, and they have recently been responsible for blooms in the Gulf of Mexico.

Toxicity

Domoic acid is structurally similar to glutamic acid, an excitatory neurotransmitter, and mimics its ability to bind to a subclass of the cognate neurotransmitter receptors at neuronal synapses [163]. Domoic acid not only binds with greater than 20 times the affinity to these receptors as the native glutamic acid [164], but it remains bound to the receptors rather than being recycled. These two characteristics allow domoic acid to create a sustained depolarization of the stimulated neuron. As a consequence, intracellular calcium levels in the stimulated neuron continually increase causing the cell to swell and burst [163]. These effects are particularly pronounced in the hippocampus, the region of brain associated with learning and memory that contains neurons that express high levels of glutamate receptor [165]. For this reason, people who consume sublethal amounts of domoic acid exhibit symptoms of amnesiac shellfish poisoning (ASP), a condition which produces short-term memory loss along with gastrointestinal sickness, vomiting, confusion/disorientation, and convulsions [166, 167]. In some patients, the memory loss lasted as long as 5 years [168].

The first, and so far only, identified ASP event that affected humans was in late 1987 around Prince Edward Island, Canada [162, 169]. Of the many people who got sick, 143 had symptoms befitting ASP and 4 of them died. It was determined that the causative agent was mussels contaminated with domoic acid [162, 167]. Subsequent analysis identified the source of the domoic acid found in the mussels as *Pseudo-nitzschia pungens* f. *multiseris* [169].

Pseudo-nitzschia produces domoic acid, and the toxin can be transferred to higher trophic levels, such as humans, through the consumption of filter-feeding bivalves that bioaccumulate domoic acid in their tissues [169–173]. Alternatively, zooplankton, krill, or planktivorous fish, e.g., anchovy, may consume toxic *Pseudo-nitzschia* spp. directly from the water column. Domoic acid may be biomagnified as secondary consumers, e.g., larger fish, birds, and marine mammals, eat the primary consumers that have digestive tracts full of toxic *Pseudo-nitzschia.* Domoic acid poisoning events have been documented in Monterey Bay that affected pelicans, cormorants, and sea lions [166, 172, 174, 175], in Mexico that affected pelicans [176], in the Gulf of California that affected loons and multiple species of marine mammals [177], and in the central California coast that affected California sea lions [178]. Because humans rarely consume whole fish, the primary and continuing concern is intoxication through consumption of contaminated shellfish or crabs.

Based on the Prince Edward Island incident and domoic acid dosage-response studies with a variety of animal models, it has been determined that acute gastrointestinal symptoms in humans develop within 24 h of ingestion followed by neurological symptoms within another 24 h. Monitoring and sampling have so far prevented additional ASP incidents in humans, so more attention has been devoted to outstanding questions regarding chronic or long-term domoic acid intoxication. Sea lions with chronic exposure to domoic acid exhibit symptoms that are distinct from those of ASP [179], which suggests similar long-term exposure could produce an additional array of detrimental effects in humans. Furthermore, domoic acid can be passed to pups in the breast milk of rodents [180]. If similar passage occurs in humans, what are the possible risks to the mental development of infants who are breast-fed from mothers who eat contaminated shellfish? Epidemiological studies are being conducted on native populations in the Pacific Northwest of the United States and Canada [181]. The diet of these native peoples relies heavily on seafood products, primarily razor clams, which are traditionally vulnerable to being toxified with varying levels of domoic acid. The results of such studies will elucidate what, if any, effects may result from consuming low levels of domoic acid over time [181].

Economics

Like other HAB events, domoic acid poisoning incidents will cause economic impacts if people and animals experience severe illness or even death, or if recreational areas and harvesting sites are closed. Beaches and shellfish-harvesting areas have faced closings in Washington state since 1991 (references within [182]). The first coastal closure in 1991 is estimated to have resulted in $23–28 million and $9 million in lost revenue to razor clam harvesters and Dungeness crab harvesters, respectively [173] (Nosho in Ref. [183]). From 1991 to 2008, about 23% of the potential razor clam fishing days along the Washington coastline have been lost due to closures from high domoic acid detected in shellfish samples (Ayers unpublished data in [184]). During an extensive domoic acid incident in 1998, the coastal tourism industry is estimated to have lost $15 million (Nosho in Ref. [183]). The first beach closure within Puget Sound occurred in 2003 [185], followed by a second closure in 2007 [182]. Ayers and Reed [186] estimate that the first closure resulted in $11.8 million in lost revenue to the local economy. The potential threat of domoic acid spreading to intracoastal regions and affecting commercial operations could amplify future losses [182]. Dyson and Huppert [184] provided a more detailed estimation of the economic effects stemming from closures of razor clam harvest sites. Based on responses of visitors and razor clam harvesters to a questionnaire, closing four beaches stretching over Grays Harbor and Pacific counties in Washington state during the 2007–2008 season would have resulted in over $25.3 million in lost expenditures. Apart from the recreational aspects of razor clam harvesting, the harvests are important to Native American cultures that rely on the razor clams and other products from the sea, not only for food but also for cultural identity.

In addition to the lost expenditures that result from closures, national, state, and local governments invest in monitoring and detection. Monitoring phytoplankton community composition, correctly identifying and enumerating *Pseudo-nitzschia* species, and measuring levels of domoic acid are all important measures for providing early warning to stakeholders located near possible blooms. Equally important is ongoing sampling of tissues from commercially and recreationally important species to monitor accumulated levels of domoic acid. The majority of methods in use for monitoring and detecting both species and toxin requires expensive equipment and can be costly to perform.

Detection

Species Identification A major hurdle for the study of *Pseudo-nitzschia* is rapid and accurate methods for species determination. Early systematics of *Pseudo-nitzschia* relied on light microscopy to separate the genus into two groups based on the width of the cells [187]. This simple categorization belied significant diversity particularly in terms of toxicity. Later, tunneling and scanning electron microscopies were employed to better resolve species complexes by analyzing such subcellular features as the striae and pore density. These methods are labor intensive, time consuming, require access to and working knowledge of electron microscopes, and still do not always adequately resolve the question of whether a sample contains toxic or nontoxic cells [188–190].

To improve detection and species identification, efforts have been made to develop molecular reagents for use in lectin-binding assays, immunoassays, sandwich hybridization, and PCR. Lectins are glycoproteins that bind sugar molecules, and lectin-binding probes have been developed that recognize species-specific glycosylated proteins on the surface of *Pseudo-nitzschia* cells [191–193]. These probes could discriminate six of seven *Pseudo-nitzschia* species endemic to New Zealand waters [192]. They showed some cross application in testing of samples from Spain and Korea [194, 195], but it was noted that the location of sample isolation and physiological state of cells can confound identification or differences between strains of the same species [196]. DNA-based probes that recognize genetic determinants such as the ribosomal RNA genes and the internal transcribed spacer (ITS) have received the most attention and have been employed in assays such as whole-cell hybridizations, FISH, sandwich hybridization, and PCR [30, 158, 159, 197–207]. Chris Scholin and his group at Monterey Bay Aquarium Research Institute have developed DNA-based probes specific for the large subunit of ribosomal RNA of *Pseudo-nitzschia* species isolated from the west coast of the United States. Similar to the above reagents, the use of these probes has met with mixed results when applied to samples at other geographic locations [203, 208–211]. For example, probes that work for *Pseudo-nitzschia pseudodelicatissima* found on the west coast of the United States did not recognize the same "species" found in the Gulf of Mexico [210], presumably due to high levels of inter- and intraspecific variability. Supporting that conclusion are results based on analyses of microsatellite markers of some *Pseudo-nitzschia* lineages or species that show high level of genetic diversity [155–159]. Additionally, one might be at fault in not identifying or recognizing the relevant markers or characteristics that determine true species differences. The Scholin probes have been further developed for use in sandwich hybridization that has been automated and incorporated into an environmental sample processor (ESP) for undersea deployment. The ESP samples seawater and quantitates both the concentration of specific harmful algal cells and their corresponding toxin concentration via the sandwich hybridization assay and automated competitive enzyme-linked immunosorbent assay, respectively [212–214]. Autonomous and continual sampling should not only provide an early warning to the presence of toxic cells but also provide important environmental data that may be used to create or enhance models for predicting the presence of such cells and their toxins.

Domoic Acid Detection and Quantitation Immediately following the Prince Edward Island incident, Canada took the lead in developing and implementing food safety standards for domoic acid. In 1989, Health and Welfare Canada instituted an action level at 20 μg domoic acid/g shellfish tissue [64]. That same level has now been adopted by the European Union, New Zealand, the United States, and Australia. One exception to this limit is the United States allows up to 30 μg/g of Dungeness crabs. If tested tissue samples exceed the limit, then the relevant beach/harvest area is closed until the domoic acid can be depurated by the organisms and repeated sampling shows that domoic acid is below the limit.

After the Prince Edward Island incident, Health and Welfare Canada relied on the same mouse bioassays as used for detecting PSTs [215]. For the reasons listed in the previous section, particularly in this case where the detection limit (∼40 μg domoic acid/g of tissue) is higher than the action level, other more sensitive and reliable methods have been developed. Some of these alternatives offer one or more advantages to the mouse bioassay but typically at the expense of time, cost, or ease of performance. Different assays also show better efficacy in analyzing domoic acid from particular sources, e.g., a receptor binding assay developed by Van Dolah et al. [216] can analyze domoic acid from shellfish and from seawater extracts of algae. A variety of antibodies have been raised and incorporated into ELISAs and competitive immunoassays for the detection of domoic acid in human body fluids or shellfish extracts [217–220]. The AOAC has a validated ELISA method (AOAC Method 2006.02) for detecting domoic acid in shellfish, and the ELISA method is used in the EU for screening purposes. A variety of modifications to a liquid chromatography (LC) methodology allow for the sensitive detection of domoic acid, but routine detection of domoic acid is often carried out using liquid chromatography coupled with diode-array/UV absorbance or fluorometic detection [221]. An LC-UV method was developed for the routine determination of domoic acid in shellfish [222]. The method was validated in a multi-lab AOAC collaborative study [223] and had an AOAC method established (AOAC Method 991.26), but a number of improvements to the extraction/clean-up procedures, alternative tandem detection systems, or other modifications have been developed and continue to be validated, e.g., [224–226]. The EU Reference Laboratory for Marine Biotoxins has created a standard operating procedure for determining or confirming domoic acid in mollusks using ultraperformance LC-MS [227]. The National Research Council of Canada was also the first to provide certified domoic acid standards for use in calibrating liquid chromatographs [97]. Similar for the detection of PSTs, a promising development is the Jellett Rapid Test or Maritime In Vitro Shellfish Test (MIST™) (Jellett Biotek Ltd.). The immunochromatographic kits can be used quickly and cheaply at sample sites in the field for the detection of domoic acid [228, 229] that must be later confirmed in the laboratory.

It is necessary to individually sample each species of interest since different species of shellfish take up and depurate domoic acid at different rates (reviewed in Ref. [163]). For example, mussels and oysters consume *Pseudo-nitzschia* at higher rates than clams, especially razor clams, but mussels and oysters also depurate much more quickly. Once domoic acid accumulates in razor clams, it can take up to a year for domoic acid levels to fall below the action level [173]. Domoic acid levels in anchovy and sand crabs rise and fall almost contemporaneously (a slight lag) with the initiation and termination of a *Pseudo-nitzschia* bloom. Scallops also depurate slowly, but the only scallop tissue that humans eat is the adductor muscle which has very low levels of domoic acid. Most domoic acid in fish, crabs, and cephalopods is in the viscera, which is typically not consumed by humans. This type of discriminatory feeding behavior is not true for animals in the wild, however, which are more likely to be affected by the presence of domoic acid moving within the food web. It is also a concern that ground whole fish is used in animal and fish feeds, and those feeds could sicken domesticated animals or fish in aquaculture far from the coasts where blooms have formed.

Origins and Nutrient Interactions

As mentioned above, *Pseudo-nitzschia* is cosmopolitan, and it is capable of taking advantage of nutrients when they are present to grow and form blooms. *Pseudo-nitzschia* has been modeled to bloom under two different scenarios of nutrient introduction. One is upwelling events during early spring and summer that bring water from the bottom up to the surface. The nutrient-rich water coupled with high turbulence from storms are generally conditions that favor diatom growth [12]. The second scenario is rivers carry natural and introduced nutrients from a variety of point and nonpoint sources to the ocean. Corresponding to the second scenario is the detection of *Pseudo-nitzschia* spp. in the Gulf of Mexico. *Pseudo-nitzschia* was considered rare in the Gulf of Mexico prior to 1950, but the numbers of several *Pseudo-nitzschia* species have increased concurrently with elevated nutrient loading into the Gulf from the Mississippi River [230]. Similar connections have been documented in the Yellow and Yangtze rivers and in the South China Sea [153, 154]. Mesocosm experiments corroborate these correlations – *Pseudo-nitzschia* show enhanced growth after pulses of nutrients are added [231].

Off the west coast of the United States, upwelling brings nutrient-rich water in contact with enormous plumes of water exiting the Puget Sound via the Strait of Juan de Fuca. A large gyre, the Juan de Fuca eddy, is formed or enhanced each spring at the junction of these two phenomena. Throughout the summer and into the fall, when it dies back, the gyre is a site of persistent upwelling that acts as a giant bioreactor for many phytoplankton species. The use of drifters to track hydrological movement has enlightened our understanding of the role of the Juan de Fuca eddy, coastal currents, and upwelling versus downwelling conditions in the growth and transport of *Pseudo-nitzschia* to various beaches and coastlines along the Pacific northwest of the United States. Modeled on the movement of drifters, *Pseudo-nitzschia* cells within the Juan de Fuca eddy are believed periodically to spin out into offshore currents. Under upwelling favorable conditions, cells will migrate southward along the Washington and Oregon state coastlines in these offshore currents. A shift to downwelling favorable conditions, such as from a storm or other events leading to a wind reversal, in regions where there is low to no river outflows will cause cells/blooms to move onshore and cause shellfish to become contaminated with domoic acid [183, 232].

The requirements for cell growth are clearly not the same as those for producing domoic acid. While the availability of nutrients is important for growth and increases in cell number, those same factors do not always coincide with domoic acid production. Different toxic species (or even the same toxic species from different geographic regions) will produce or not produce domoic acid in response to different environmental conditions (including the availability of one or more limiting nutrients such as phosphorus or silicate), stages of culture growth (e.g., exponential vs stationary growth phases), and the composition of the associated bacterial community [161, 233–237]. It has also been suggested that iron limitation or copper stress can induce domoic acid to act as either a chelator of iron (to enhance its uptake into the cell) or copper (to alleviate its toxicity) [238]. Iron stimulation experiments in the Southern Ocean show increased *Pseudo-nitzschia* cell abundance, but not a concomitant increase in domoic acid [239, 240]. Axenic cultures of *Pseudo-nitzschia* produce less domoic acid than nonaxenic cultures [241–243] indicating a role for bacteria, although no bacterium has been shown capable of producing domoic acid on its own.

Cyanobacteria

Description

Cyanobacteria, originally classified as "blue-green algae," are prokaryotes capable of photosynthesis (reviewed in Ref. [244]). They underwent a worldwide adaptive radiation that has allowed them to fill a wide range of ecological niches including terrestrial and aquatic systems (both freshwater and marine) as well as forming numerous endo- and ectosymbiotic associations [245]. Their distribution and diversity are likely a function of their ancient origins – it is estimated that they have existed for at least 2 billion years and possibly as long as 3.5 billion years or more. They may have been one of the first forms of life on Earth (reviewed in Ref. [246]). For this reason, the diversity of cyanobacterial species is also reflected in the wide range of metabolites and secondary molecules that are produced within this phylum. Some of these unique molecules act as toxins, carcinogens, and/or allergens when ingested, inhaled, or contacted by humans and animals. Coupled with the ability of some fresh and marine water species to form blooms, cyanotoxins pose a wide range of threats to human health.

Toxicity

The first recorded deaths from cyanotoxins occurred in Australia in 1878 when livestock drank from a brackish, estuarine lake covered with a *Nodularia spumigena* bloom [247]. Since then, there have been many other reported instances of sickness or death of livestock, pets, and humans attributed to recreational exposure to or ingestion of cyanobacteria-contaminated water (reviewed in Ref. [248–253]). A comprehensive description, or even listing, of all identified or suspected toxic molecules from cyanobacteria is beyond the scope of this article. Listed in Table 1 are a few representative toxins from cyanobacteria.

Hepatotoxins Microcystins are cyclic heptapeptides that include nonprotein amino acids and unique covalent linkages. They are the most frequently found cyanotoxin, and over 70 variants have been described [254]. They are highly stable and water-soluble. Ingestion of microcystins can lead to a rapid symptomatology (within 30 min to 24 h) that may ultimately lead to hemorrhaging in the liver and death [255, 256]. It is also possible that inhalation of microcystins (or other cyanotoxins) could be a route of exposure since microcystins have been detected, albeit at low levels, in aerosol droplets [257]. A major research focus for microcystins, indeed for all the cyanotoxins, is determining what effects low level or chronic exposure may produce. Mounting evidence suggests that microcystins promote tumor growth and certain cancers [258–262]. The World Health Organization has set a provisional consumption limit (for drinking water) of 1 μg/L for the most toxic form of microcystin [263].

Harmful Algal Blooms. Table 1 Representative toxins listed with their genera of origin and toxicity

Toxin	Representative genera[a]	Ecosystem	Toxicity
Microcystins	*Microcystis, Anabaena, Nostoc, Planktothrix*	Fresh water	Hepatotoxin
Nodularins	*Nodularia*	Brackish water	Hepatotoxin
		Marine water	
Cylindrospermopsin	*Cylindrospermopsis, Anabaena*	Fresh water	Hepatotoxin
Anatoxin-a	*Anabaena, Planktothrix*	Fresh water	Neurotoxin
Anatoxin-a(s)	*Anabaena*	Fresh water	Neurotoxin
Saxitoxins	*Anabaena*	Fresh water	Neurotoxin
	Lyngbya	Marine water	
	Cylindrospermopsins	Fresh water	
	Trichodesmium	Marine water	
LPS	All cyanobacteria	All	Endotoxin, Dermatoxin
Aplysiatoxin	*Lyngbya*	Marine water	Dermatoxin, Tumor promoter
	Planktothrix	Fresh water	
Lyngbyatoxin-a	*Lyngbya*	Marine water	Dermatoxin

[a] Not all species within the listed genera necessarily produce the associated toxin, and species in other genera not listed may also produce these toxins

A real concern is the fact that "finished" water from some water treatment facilities has tested above that limit (reviewed in Ref. [264–266]). Microcystin levels can be greater than 24 mg/L in waterbodies with very high cell counts [267], which can pose risks to recreational swimmers based on an average accidental ingestion of 100–200 mL of water. Freshwater-derived microcystins have also been found in marine shellfish located at sites where rivers discharge into marine waters, and these toxins are the implicated source for causing hepatotoxic shellfish poisoning (HSP) that led to the death of 21 southern sea otters in Monterey Bay, California [268]. Microcystins can be bioaccumulated to high levels in the tissue of shellfish that, when coupled with a slow depuration rate [268], poses additional health risks to shellfish consumers, including humans.

To date, nodularins are known to be produced by the single species, *Nodularia spumigena*, a species of cyanobacteria that can form blooms in brackish and marine waters. Nodularins are structurally similar to microcystins (cyclic pentapeptides), and their toxicity and mode of action are similar as well. One distinction between the two classes of hepatotoxins, however, is that nodularins are carcinogenic, i.e., they can initiate tumor formation, as well as act as potent tumor promoters [254].

Cylindrospermopsins primarily produce effects in the liver, but they can also damage other organs. These toxins are guanidine alkaloids that take days, rather than hours, to produce symptoms of intoxication. Additional toxic effects include the inhibition of protein synthesis and genotoxicity [269].

Neurotoxins Saxitoxins have been discussed in a previous section of this article (see "Toxicity" under "*Alexandrium*"). Anatoxin-a and anatoxin-a(s) are structurally unrelated, but both can act swiftly (within minutes to hours) to disrupt the function at nerve synapse leading to muscle paralysis and respiratory arrest [270]. Anatoxin-a can be inhaled, ingested, or absorbed through the skin. The estimated toxic dose for a human adult male is less than 5 mg. Anatoxin-a(s) has only been found in two species of *Anabaena, A. flos-aque* and *A. lemmermannii* [271, 272]. It is the only known natural organophosphate. As such, it acts by inhibiting acetylcholinesterase, and it is ten times more potent than anatoxin-a [272, 273]. Fortunately, the occurrence and concentration of these two toxic molecules are much less than are typically detected for microcystins. While humans will not likely ingest enough cells/neurotoxin(s) in a mouthful of water, some animals that may drink liters at a time may be at risk.

Other Cyanobacterial Toxins Lipopolysaccharides (LPS) are components of the outer membrane of gram-negative bacteria and cyanobacteria. They are capable of forming complexes with several biomolecules to elicit irritant and allergic responses [274]. Cyanobacterial LPS can cause skin irritation, gastroenteritis, and fever, or act as an endotoxin that can kill mice in mouse bioassays or exacerbate the effects of other cyanotoxins [275].

Aplysiatoxin and lyngbyatoxin-a are only produced in marine species of cyanobacteria. Generally, the concern surrounding toxins from marine cyanobacteria is not ingestion from drinking water but either trophic transfer and biomagnifications in fish, shellfish, and other consumables or contact exposure/inhalation from recreating around blooms. Aplysiatoxin and lyngbyatoxin-a can cause severe dermatitis through acute inflammatory reactions [276, 277]. They can also activate protein kinase C, a serine/threonine kinase that is central to diverse signal transduction pathways, to potently promote tumor formation [278, 279]. Sea turtles that consume seaweed overgrown or associated with *Lyngbya*, a genus of marine cyanobacteria, are suspected to bioaccumulate lyngbyatoxins in their tissues, and there has been one potential human death, and other suspected cases, linked to the consumption of sea turtle meat contaminated with lyngbyatoxins [280, 281].

β-*N*-Methylamino-L-Alanine (BMAA) In addition to any of the acute effects listed above, long-term chronic effects, e.g., cancer development, are linked or suspected to several cyanotoxins [258–262]. Additional cyanobacterial metabolites that are receiving attention include unusual amino acids such β-*N*-methylamino-L-alanine (BMAA). BMAA has been proposed as a causative agent of neurotangle diseases, e.g., Parkinson's disease, amyotrophic lateral sclerosis, or Alzheimer's disease. Such a role is controversial, but it is supported by a number of correlative studies including direct detection of BMAA in brain tissue samples from native individuals from the island of Guam who died from lytico-bodig, a type of neurotangle disease, and a small sampling of Caucasian patients in Canada who died from Alzheimer's disease (reviewed in Ref. [282]). In vitro and in vivo laboratory studies using BMAA support the purported role as a neurotoxin in that BMAA leads to selective motor neuron death, produces behavioral and/or neurological differences, and interferes with learning and memory [283–288].

A second, similar biomagnification mechanism may be responsible for the occurrence of avian vacuolar myelinopathy (AVM) in aquatic and predatory birds. In the southeastern United States, *Hydrilla verticillata*, is an aquatic plant that often contains a BMAA-producing epiphytic cyanobacteria [289, 290]. Herbivorous birds, e.g., ducks and coots, which consume *H. verticillata* (and bald eagles that consume the herbivorous birds), contract high rates of AVM and die [290]. Although the majority of the evidence for vectoring of BMAA to humans and birds is correlative, some direct tests of the hypotheses support the proposed mechanisms (e.g., [291–293]). Also of note, and concern, is the fact that BMAA has been detected in almost all cyanobacteria in all environments worldwide (e.g., [294–298]). Intriguingly, a recent epidemiological study implicates a lake, noted for frequent cyanoblooms, as being responsible for or linked to a 10–25× higher incidence of sporadic ALS, compared to the general US population, in lake-adjacent areas [299].

Economics

In addition to the public health consequences, additional economic and sociocultural impacts caused by cyanoblooms include negative effects to: (1) recreation and tourism due to closure of waterbodies (often at times when recreational use is most desired or when tourism would be at peak, i.e., summer months); (2) aquaculture due to loss of product (death) or production of unwanted, tainted product (some cyanobacterial metabolites can alter the color or taste of the fish) [275]; and (3) the quality of life of people living and working around affected waterbodies due to the reduced aesthetic appeal and/or the unwanted health risks. Few studies have quantified the losses or impacts of cyanobacterial blooms. Tucker et al. [300] estimated that greater than $80 million was lost per year in catfish aquaculture facilities due to cyanobacterial metabolites that produce off-color or off-flavor in the catfish. In addition to toxic effects, cyanobacteria can form nuisance blooms that generate costs associated with cleaning biomass that blocks water intakes used for draining and industrial purposes such as hydroelectric power generation. A conservative estimate of economic costs of all freshwater HABs in the Unites States is $2.2–4.6 billion a year [301].

Where species of known or potential bloom formers are present, monitoring is essential as both a tool for aiding in possible prevention of blooms and to alert the public of existing or impending cyanoblooms. Public guidance should include precautions about avoiding areas of visible algal biomass, about avoiding direct contact with blooms, and to avoid swallowing water containing algae as those interactions are associated the highest risks to human/animal health. A bloom may reach densities of >100,000 cells/mL, but even at lower levels, the concentration of toxic metabolites may be sufficient to cause sickness or death. The WHO had published recommendations for recreational exposure, based in part upon results from [249, 255]: Up to 20,000 cyanobacterial cells/mL requires a warning to visitors but has a low/mild risk to adverse health effects, up to 100,000 cells/mL carries a moderate risk in the short-term as well as potential long-term risks if swallowed, and when a visible bloom forms, then a high risk of inhalation, ingestion, or body contact is likely at that site [302]. Monitoring and public alerts about the presence and dangers of cyanoblooms are recurring costs that state and local agencies will be forced to carry. Atech (2000) (in Ref. [303]) estimated costs related to monitoring for cyanobacteria in Australia at greater than $12 million dollars per year (in 2011 US dollars).

Another significant expense is the detection and removal of cyanotoxins from drinking water sources. It is axiomatic that preventing blooms and/or toxin production should be cheaper, easier, and safer than eliminating or removing cells and/or toxins from the water after a bloom. The primary and essential goal of cyanoHAB management, if not elimination, is identifying the type and source of nutrient loading in order to reduce cyanobacterial proliferation. To remove cells once a bloom has formed, however, has typically relied on common algaecides such as copper sulfate. Unfortunately, the resulting cell lysis can exacerbate the problems as intracellular cyanotoxins are released into the water, increasing their concentration. Additionally, the indiscriminate or repetitive use of copper-based reagents can cause inadvertent ecological damage and induce copper resistance in surviving algae [304–306]. Other methods for bloom control include destratification, which only works for small waterbodies, and increased flushing, which only works if upstream sources of water are available. Conventional water treatments such as flocculation, coagulation, chlorination, and filtration can remove cells; but not only do they not remove toxins, they also cause cell lysis leading to increased toxin levels in the water [307–312]. While some methods of oxidation, such as chlorination, do not appear effective [307–312], others such as UV disinfection, ozonation, and potassium permanganate treatment are more promising [265, 313]. Pipelines of multiple methodologies, e.g., use of activated charcoal (particularly wood-based, powdered activated carbon), oxidation via ozonation, and nanofiltration (with nominal molecular weight cutoff of 200) appear effective at removing microcystins and anatoxin-a [265, 313]. Titanium dioxide photocatalysis is being developed for both purification and disinfection. This method can destroy microcystins found in concentrations up to 5,000 μg/L [8, 314]. Not all of the methods listed above are universally employed and are absolutely lacking in sources of water such as private wells. Also, these treatments do not protect wildlife from consuming water at natural waterbodies where cyanoblooms may be occurring, all of which argue for preventive management of cyanoblooms rather than reactive management.

Detection

Microscopic inspection of water samples is still primarily used for identification and quantitation of cyanobacterial cells. Assaying for chlorophyll-*a* is a quick alternative to microscopy but may overestimate the amount of cyanobacteria in a sample because chlorophyll-*a* is also present in many other algae [255]. It is also a presumptive indicator of cyanotoxins rather that a direct detection method. Immunoassays [315] following standard practices outlined in Chorus and Bartram [255] have also been developed for these purposes. More recently, genetic identification using polymerase chain reaction and modifications thereof, e.g., quantitative polymerase chain reaction, are being used to detect (and quantify) specific species or toxin-synthesizing genes, e.g., a nodularin synthetase gene in *Nodularia* [316] or a microcystin synthetase gene in *Microcystis* [317], within an environmental sample.

For the detection of cyanotoxins, the mouse bioassay [256] is the official approved method, but a multitude of more-advanced, sensitive, and humane methodologies have been and continue to be developed. Some methods have been adapted to detect more than one type of cyanotoxin, while other methods are specific for one class or variant, e.g., in vitro biological assays such as acetylcholine esterase inhibition (to detect anatoxin-a(s)), protein phosphatase inhibition (to detect microcystins or nodularins), and protein synthesis inhibition (to detect cylindrospermopsins) [318]. Modifications to the standard high-pressure liquid chromatography methodology, e.g., UV detection at the respective absorption maximum of each cyanotoxin, following separation allows for the detection of microcystins, cylindrospermopsins, and anatoxin-a [254]. Liquid chromatography (LC) with post-column oxidation and fluorescence detection is now an official method of analysis by the AOAC for saxitoxins [98]. Similar coupling of LC with tandem mass spectrometry is also useful for detecting anatoxin-a(s) [319]. Antibodies against microcystin, nodularin, saxitoxin, and cylindrospermopsin have been useful for developing ELISAs [72, 320, 321] that show good correlation with HPLC methods. The detection methods have greatly benefited from the creation and commercialization of kits and cyanotoxin standards. Standard hemoagglutination assays, e.g., the Limulus amoebocyte lysate assay, can be used to quantify LPS in aqueous samples [322].

Origins and Nutrient Interactions

Cyanobacteria are ubiquitous [323, 324], and the threat of cyanoblooms is pervasive. Similar to other HAB species, the frequency and geographic distribution of cyanoHABs is increasing worldwide [8, 325]. A 1999–2000 survey in Florida sampled 75 bodies of water and

88/167 samples had cyanotoxins [326]. Eighty percent of the positive samples had levels sufficient to kill a mouse in a mouse bioassay [326]. In a survey of 45 water sources in the United States and Canada between 1996 and 1998, 80% had detectable levels of microcystin – a few above the WHO advisory limit [266]. Other surveys of national or international scope indicate that about 60% of bodies of fresh water contain toxic species of cyanobacteria [266, 297, 327]. The North American Great Lakes and associated water systems are a source of drinking water for ∼22 million people, and during late summer months, microcystin sampling routinely exceeds the WHO threshold [328]. Over the long term, e.g., a 75-year lifetime, it has been estimated using the data from Carmichael [266] that cyanotoxin exposure from drinking water is an order of magnitude below the WHO guideline of 1 μg/L [329]. Such a study is somewhat reassuring, but the effects of low-level chronic exposure are still not understood; nor can it be expected that cyanotoxin levels in finished drinking water will not increase unless measures are taken to decrease nutrient loading to decrease bloom formation or methods to more effectively remove cyanotoxins from the water are developed.

Some cyanobacterial blooms occur naturally under specific conditions, but the frequencies are much higher in eutrophic bodies of water [330]. Other significant factors that support bloom formation are warmer temperatures, increased sunlight, and decreased flow/flushing of a waterbody [305]. The induction of bloom formation in stagnant waterbodies or slow-moving, low-flushing systems is clearly linked to nutrient loading, particularly phosphorous [331]. In such systems, gas-filled vesicles allow cyanobacteria to be buoyant and remain suspended near the surface to be exposed to light [332]. Even under conditions of limiting nitrogen or phosphorous, cyanobacteria can outcompete other planktonic organisms due to higher affinities for these nutrients and the capacity to store high amounts of phosphorous inside their cells [333]. In waterbodies that are mixed and/or flushed, factors such as warmer temperatures, high amounts of sunlight, and elevated nutrients will all contribute to different degrees in leading to cyanobloom formation (reviewed in Ref. [255, 333]). The convergence of the optimal conditions for a particular cyanobacterial species will allow it to form a "scum" or surface bloom. Where applicable or possible, efforts by watershed managers to reduce excess nutrient loading, to increase or maintain water flow/flushing rates, and to prevent the removal of grazers have all been effective strategies to reduce cyanoblooms (reviewed in Ref. [305]). The introduction of solar-powered circulation (SPC) devices into nutrient-enriched, cyanobacteria-containing reservoirs has been effective at suppressing the number of cyanobacterial cells and decreasing the need for supplemental copper sulfate (algaecide) applications [334]. Broad-based deployment of such devices would not only decrease the occurrence of cyanoHABs but also do so in an ecologically and environmentally sound manner.

Future Directions

Significant effort has generated a lot of biological and ecological knowledge of selected harmful algae. This knowledge has produced few methods to prevent bloom formation or mitigate the effects of blooms that have formed. Few attempts to control HABs have proven to be practical or free of adverse side effects that would further disrupt ecosystems. The difficulty in preventing HABs is that each species reacts to a different suite of factors; some of these factors can include a complex mixture of biochemical, ecological, geophysical, hydrological, and meteorological conditions. In the majority of cases, any of these conditions may be beyond our control except for one. The one almost universal requirement for any algal blooms is a supply of nutrients. Without nutrients, the likelihood of any species forming a bloom is low, so identifying the source(s) of relevant nutrients and then minimizing those nutrients may be the best strategy for preventing and mitigating harmful algal blooms.

The most advances in HAB research have been in the ability to detect/monitor HAB species and toxins. Despite these advances, continued development of appropriate standards and techniques for regulatory and management practices is imperative, especially because of the continuing discovery of new harmful species. An emerging line of research to aid in detection and modeling cell/bloom behavior is at the molecular level. A lot of new research is beginning to elucidate the genetic composition of these different species and the mechanisms that regulate the genetic responses, and subsequent physiological responses, to a variety of environmental parameters. These data sets will prove invaluable to refine or develop current or future models, respectfully, that will be useful for predicting bloom formation and behavior.

Bibliography

Primary Literature

1. Hallegraeff GM (1993) A review of harmful algal blooms and their apparent global increase. Phycologia 32:79–99
2. Sandifer P et al (2007) Interagency oceans and human health research implementation plan: a prescription for the future. Joint Subcommittee on Ocean Science and Technology. Interagency Working Group on Harmful Algal Blooms, Hypoxia and Human Health, Washington, DC
3. Smayda TJ (1990) Novel and nuisance phytoplankton blooms in the sea: evidence for a global epidemic. In: Graneli E (ed) Toxic marine phytoplankton. Elsevier, New York, pp 29–40
4. Smayda TJ (1992) Global epidemic of noxious phytoplankton blooms in the sea: evidence for a global epidemic. In: Sherman K, Alexander LM, Gold BD (eds) Food chains: models and management of large marine ecosystems. Westview Press, San Francisco, pp 275–307
5. Smayda TJ, White AW (1990) Has there been a global expansion of algal blooms? If so is there a connection with human activities? In: Graneli E (ed) Toxic marine phytoplankton. Elsevier, New York, pp 516–517
6. Hallegraeff GM (2010) Ocean climate change, phytoplankton community responses, and harmful algal blooms: a formidable predictive challenge. J Phycol 46(2):220–235
7. Ramsdell JS, Anderson DM, Glibert PM (eds) (2005) HARRNESS, Harmful algal research and response: a national environmental science strategy 2005–2015. Ecological Society of America, Washington, DC
8. Lopez CB et al (2008) Scientific assessment of marine harmful algal blooms. Joint Subcommittee on Ocean Science and Technology. Interagency Working Group on Harmful Algal Blooms, Hypoxia and Human Health, Washington, DC
9. Heisler J et al (2008) Eutrophication and harmful algal blooms: a scientific consensus. Harmful Algae 8(1):3–13
10. Anderson DM et al (2008) Harmful algal blooms and eutrophication: examining linkages from selected coastal regions of the United States. Harmful Algae 8(1):39–53
11. Huisman JM, Saunders GW (2007) Phylogeny and classification of the algae. In: McCarthy PM, Orchard AE (eds) Algae of Australia: introduction. Australian Biological Resources Study/CSIRO, Melbourne, pp 66–103
12. Smayda TJ, Reynolds CS (2001) Community assembly in marine phytoplankton: application of recent models to harmful dinoflagellate blooms. J Plankton Res 23(5):447–461
13. Vargo GA et al (2001) The hydrographic regime, nutrient requirements, and transport of a *Gymnodinium breve* Davis red tide on the west Florida shelf. In: Hallegraeff GM et al (eds) Harmful algal blooms 2000. Intergovernmental Oceanographic Commission of UNESCO, Paris, pp 157–159
14. Laws RM (1985) The ecology of the southern ocean. Am Sci 73(1):26–40
15. Croxall JP, Nicol S (2004) Management of southern ocean fisheries: global forces and future sustainability. Antarct Sci 16(4):559–584
16. Smayda TJ, Reynolds CS (2003) Strategies of marine dinoflagellate survival and some rules of assembly. J Sea Res 49(2):95–106
17. Sournia A (1974) Circadian periodicities in natural populations of marine phytoplankton. Mar Biol 12:325–389
18. Kent ML, Whyte JNC, Latrace C (1995) Gill lesions and mortality in seawater pen-reared Atlantic salmon Salmo-Salar associated with dense bloom of *Skeletonema costatum* and *Thalassiora* species. Dis Aquat Organ 22(1):77–81
19. Bell GR (1961) Penetration of spines from a marine diatom into gill tissue of Lingcod (*Ophiodon elongatus*). Nature 192(479):279–280
20. Dagg MJ et al (2007) A review of water column processes influencing hypoxia in the northern Gulf of Mexico. Estuaries Coast 30(5):735–752
21. Rabalais NN et al (2002) Nutrient-enhanced productivity in the northern Gulf of Mexico: past, present and future. Hydrobiologia 475(1):39–63
22. Legrand C et al (2003) Allelopathy in phytoplankton – biochemical, ecological, and evolutionary aspects. Phycologia 42:406–419
23. Kubanek J et al (2005) Does the red tide dinoflagellate *Karenia brevis* use allelopathy to outcompete other phytoplankton? Limnol Oceanogr 50(3):883–895
24. Okaichi T, Nishio S (1976) Identification of ammonia as the toxic principle of red tide of *Noctiluca miliaris*. Bull Plankton Soc Jpn 23:75–80
25. Lancelot C et al (2011) Cost assessment and ecological effectiveness of nutrient reduction options for mitigating Phaeocystis colony blooms in the southern North Sea: an integrated modeling approach. Sci Total Environ 409(11):2179–2191
26. Ciglenecki I et al (2003) Mucopolysaccharide transformation by sulfide in diatom cultures and natural mucilage. Mar Ecol Prog Ser 263:17–27
27. Guiry MD (2011) AlgaeBase. http://www.algaebase.org [cited 10 May 2011]
28. Daugbjerg N et al (2000) Phylogeny of some of the major genera of dinoflagellates based on ultrastructure and partial LSU rDNA sequence data, including the erection of three new genera of unarmoured dinoflagellates. Phycologia 39:302–317
29. Scholin CA, Anderson DM (1994) Identification of group-specific and strain-specific genetic-markers for globally distributed *Alexandrium* (Dinophyceae). 1. RFLP Analysis of SSU ribosomal-RNA genes. J Phycol 30(4):744–754
30. Scholin CA, Anderson DM (1996) LSU rDNA-based RFLP assays for discriminating species and strains of *Alexandrium* (*Dinophyceae*). J Phycol 32(6):1022–1035

31. Scholin CA, Hallegraeff GM, Anderson DM (1995) Molecular evolution of the *Alexandrium tamarense* species complex (Dinophyceae)-dispersal in the North American and West Pacific regions. Phycologia 34:472–485
32. Scholin CA et al (1994) Identification of group-specific and strain-specific genetic-markers for globally distributed *Alexandrium* (Dinophyceae) 2 sequence-analysis of a fragment of the Lsu ribosomal-RNA gene. J Phycol 30(6):999–1011
33. Lilly EL, Halanych KM, Anderson DM (2007) Species boundaries and global biogeography of the *Alexandrium tamarense* complex (Dinophyceae). J Phycol 43:1329–1338
34. Kalaitzis JA et al (2010) Biosynthesis of toxic naturally-occurring seafood contaminants. Toxicon 56:244–258
35. Etheridge SM (2010) Paralytic shellfish poisoning: seafood safety and human health perspectives. Toxicon 56: 108–122
36. Oshima Y, Blackburn SI, Hallegraeff GM (1993) Comparative study on paralytic shellfish toxin profiles of the dinoflagellate *Gymnodinium catenatum* from three different countries. Mar Biol 116:471–476
37. Hall S et al (1990) The saxitoxins: sources, chemistry, and pharmacology. In: Hall S, Strichartz G (eds) Marine toxins: origins, structure, and molecular pharmacology. American Chemical Society, Washington, DC
38. Hille B (1992) Ionic channels of excitable membranes, 2nd edn. Sinauer, Sunderland
39. Ritchie JM, Rogart RB (1977) The binding of saxitoxin and tetrodotoxin to excitable tissue. Rev Physiol Biochem Pharmacol 79:42–50
40. Kao CY et al (1967) Vasomotor and respiratory depressant actions of tetrodotoxin and saxitoxin. Arch Int Pharmacodyn Ther 165:438–450
41. Kao CY (1993) Paralytic shellfish poisoning. In: Falconer IR (ed) Algal toxins in seafood and drinking water. Academic, London, pp 75–86
42. Doucette G et al (2006) Phycotoxin pathways in aquatic food webs: transfer, accumulation, and degradation. In: Graneli E, Turner JT (eds) Ecology of harmful algae. Springer, Berlin/ Heidelberg
43. Anderson DM, White AW (1989) Toxic dinoflagellates and marine mammal mortalities. Woods Hole Oceanographic Institution Technical Report. WHOI-89-36 (CRC-89-6). Woods Hole Oceanographic Institution, Woods Hole
44. Geraci JR (1989) Clinical investigations of the 1987–1988 mass mortality of bottlenose dolphins along the US central and south Atlantic coast. In: Final report to the national marine fisheries service. U. S. Navy, Office of Naval Research and Marine Mammal Commission, Ontario Veterinary College, University of Guelph, Guelph, pp 1–63
45. Geraci JR et al (1989) Humpback whales (*Megaptera novaeangliae*) fatally poisoned by dinoflagellate toxin. Can J Fish Aquat Sci 46:1895–1898
46. Shumway SE (1995) Phycotoxin-related shellfish poisoning: bivalve molluscs are not the only vectors. Rev Fish Sci 3:1–31
47. White AW (1981) Marine zooplankton can accumulate and retain dinoflagellate toxins and cause fish kills. Limnol Oceanogr 26:103–109
48. Doucette GJ et al (2006) PSP toxins in North Atlantic right whales (*Eubalaena glacialis*) and their zooplankton prey in the Bay of Fundy. Canada Mar Ecol Prog Ser 306:303–313
49. Deeds JR et al (2008) Non-traditional vectors for paralytic shellfish poisoning. Mar Drugs 6:308–348
50. Shumway SE (1990) A review of the effects of algal blooms on shellfish and aquaculture. J World Aquacul Soc 21:65–104
51. RaLonde R (1996) Paralytic shellfish poisoning: the Alaska problem. In: Alaska's marine resources. Marine Advisory Program, Anchorage
52. Bricelj VM, Lee JH, Cembella AD (1991) Influence of dinoflagellate cell toxicity on uptake and loss of paralytic shellfish toxins in the northern quahog *Mercenaria mercenaria*. Mar Ecol Prog Ser 74:33–46
53. Reboreda A et al (2010) Decrease in marine toxin content in bivalves by industrial processes. Toxicon 55:235–243
54. Trainer VL (2002) Harmful algal blooms on the U.S. west coast. In: Taylor FJ, Trainer VL (eds) Harmful algal blooms in the PICES region of the North Pacific. PICES Scientific Report No. 23. North Pacific Marine Science Organization, Sidney, pp 89–118
55. Nishitani L, Chew KK (1988) PSP toxins in the Pacific coast states: monitoring programs and effects on bivalve industries. J Shellfish Res 7:653–669
56. Anderson DM et al (2000) Estimated annual economic impacts from harmful algal blooms (HABs) in the US. Woods Hole Oceanographic Institute, Woods Hole
57. Anderson DM et al (2005) Initial observation of the 2005 *Alexandrium fundyense* bloom in southern New England: general patterns and mechanisms. Deep Sea Res II 52:2856–2876
58. Jin D, Thunberg E, Hoagland P (2008) Economic impact of the 2005 red tide event on commercial shellfish fisheries in New England. Ocean Coast Manag 51(5):420–429
59. Adachi M, Sako Y, Ishida Y (1996) Analysis of *Alexandrium* (Dinophyceae) species using sequence of the 5.8S ribosomal DNA and internal transcribed spacer regions. J Phycol 32:424–432
60. John U, Fensome RA, Medlin LK (2003) The application of a molecular clock based on molecular sequences and fossil record to explain biogeographic distributions within the *Alexandrium tamarense* "species complex" (Dinophyceae). Mol Biol Evol 20:1015–1027
61. Medlin L et al (1998) Sequence comparison links toxic European isolates of *Alexandrium tamarense* from the Orkney Islands to toxic North American stocks. Eur J Protistol 34:329–335
62. Hollingworth T, Wekell MM (1990) Paralytic shellfish poison biological method. Final action. In: Hellrich K (ed) Official methods of analysis. AOAC, Arlington, pp 881–882
63. AOAC (2000) AOAC official method 959.08 paralytic shellfish poison. In: Horwitz W (ed) Official methods of analysis of AOAC international. AOAC International, Arlington

64. Wekell JC, Hurst J, Lefebvre KA (2004) The origin of the regulatory limits for PSP and ASP toxins in shellfish. J Shellfish Res 23:927–930
65. Food and Agriculture Organization of the United Nations (2004) Marine biotoxins. Paper 80. FAO Food and Nutrition, Rome, pp 5–52
66. Jellet JF et al (1992) Paralytic shellfish poison (saxitoxin family) bioassays: automated endpoint determination and standardization of the in vitro tissue culture bioassay and comparison with the standard mouse bioassay. Toxicon 30: 1143–1156
67. Kogure K et al (1988) A tissue culture assay for the tetrodotoxin, saxitoxin, and related toxins. Toxicon 26:191–197
68. Shimojo RY, Iwaoka WT (2000) A rapid hemolysis assay for the detection of sodium channel-specific marine toxins. Toxicology 154:1–7
69. Davio SR, Fontelo PA (1984) A competitive displacement assay to detect saxitoxin and tetrodotoxin. Anal Biochem 141:199–204
70. Doucette GJ et al (1997) Development and preliminary validation of a microtiter plate-based receptor binding assay for paralytic shellfish poisoning toxins. Toxicon 35(5):625–636
71. Powell CL, Doucette GJ (1999) A receptor binding assay for paralytic shellfish poisoning toxins: recent advances and applications. Nat Toxins 7:393–400
72. Chu FS, Huang X, Wei RD (1990) Enzyme-linked immunosorbent assay for microcystins in blue-green algal blooms. J Assoc Off Anal Chem 73(3):451–456
73. Usleber E, Schnieder E, Terplan G (1991) Direct enzyme immunoassay in microtitration plate and test strip format for the detection of saxitoxin in shellfish. Lett Appl Microbiol 13:275
74. Bergantin JH, Sevilla F III (2010) Quartz crystal microbalance biosensor for saxitoxin based on immobilized sodium channel receptors. Anal Lett 43(3):476–486
75. Metfies K et al (2005) Electrochemical detection of the toxic dinoflagellate *Alexandrium ostenfeldii* with a DNA-biosensor. Biosens Bioelectron 20(7):1349–1357
76. Diercks S, Metfies K, Medlin LK (2008) Development and adaptation of a multiprobe biosensor for the use in a semi-automated device for the detection of toxic algae. Biosens Bioelectron 23:1527–1533
77. Fonfria ES et al (2007) Paralytic shellfish poisoning detection by surface plasmon resonance-based biosensors in shellfish matrixes. Anal Chem 79(16):6303–6311
78. Chu FS et al (1996) Screening of paralytic shellfish poisoning toxins in naturally occurring samples with three different direct competitive enzyme-linked immunosorbent assays. J Agric Food Chem 44(12):4043–4047
79. Garthwaite I et al (2001) Integrated enzyme-linked immunosorbent assay screening system for amnesic, neurotoxic, diarrhetic, and paralytic shellfish poisoning toxins found in New Zealand. J AOAC Int 84:1643–1648
80. Kawatsu K et al (2002) Development and application of an enzyme immunoassay based on a monoclonal antibody against gonyautoxin components of paralytic shellfish poisoning toxins. J Food Prot 65(8):1304–1308
81. Usleber E et al (2001) Immunoassay methods for paralytic shellfish poisoning toxins. J AOAC Int 84:1649–1656
82. Usleber E, Donald M, Straka M (1997) Comparison of enzyme immunoassay and mouse bioassay for determining paralytic shellfish poisoning toxins in shellfish. Food Addit Contam 14(2):193–198
83. Gas F et al (2010) One step immunochromatographic assay for the rapid detection of *Alexandrium minutum*. Biosens Bioelectron 25:1235–1239
84. Jellet JF et al (2002) Detection of paralytic shellfish poisoning (PSP) toxins in shellfish tissue using MIST Alert, a new rapid test, in parallel with the regulatory AOAC mouse bioassay. Toxicon 40(10):1407–1425
85. Costa PR et al (2009) Comparative determination of paralytic shellfish toxins (PSTs) using five different toxin detection methods in shellfish species collected in the Aleutian islands, Alaska. Toxicon 54:313–320
86. Oshiro M et al (2006) Paralytic shellfish poisoning surveillance in California using the Jellett Rapid PSP test. Harmful Algae 5:69–73
87. Sullivan JJ (1988) Methods of analysis for DSP and PSP toxins in shellfish: a review. J Shellfish Res 7:587–595
88. Dell'Aversano C, Hess P, Quilliam MA (2005) Hydrophilic interaction liquid chromatography-mass spectrometry analysis of paralytic shellfish poisoning (PSP) toxins. J Chromatogr A1081:190
89. Oshima Y (1995) Post-column derivation HPLC methods for paralytic shellfish poisons. In: Hallegraeff GM, Anderson DM, Cembella AD (eds) Manual on harmful marine microalgae. IOC of UNESCO, Paris, pp 81–94
90. Rourke WA et al (2008) Rapid postcolumn methodology for determination of paralytic shellfish toxins in shellfish tissue. J AOAC Int 91(3):589–597
91. Thomas K et al (2006) Analysis of PSP toxins by liquid chromatography with postcolumn oxidation and fluorescence detection. In: Henshilwood K et al (eds) Molluscan shellfish safety. The Marine Institute, Galway, pp 132–138
92. Lawrence JF, Menard C (1991) Liquid chromatographic determination of paralytic shellfish poisons in shellfish after prechromatographic oxidation. J Assoc Off Anal Chem 74:1006–1012
93. Lawrence JF et al (1991) A study of ten toxins associated with paralytic shellfish poison using prechromatographic oxidation and liquid chromatography with fluorescence detection. J Assoc Off Anal Chem 74:419–425
94. Lawrence JF, Niedzwiadek B, Menard C (2004) Quantitative determination of paralytic shellfish poisoning toxins in shellfish using prechromatographic oxidation and liquid chromatography with fluorescence detection: interlaboratory study. J AOAC Int 87(1):83–100
95. Quilliam MA, Janecek M, Lawerence JF (1993) Characterization of the oxidation products of paralytic shellfish poisoning toxins by liquid chromatography/mass spectrometry. Rapid Commun Mass Spectrom 6:14–24
96. Laycock MV et al (1994) Isolation and purification procedures for the preparation of paralytic shellfish poisoning toxin standards. Nat Toxins 2:175–183

97. NRC (2003) National Research Council Canada (NRC) – Institute for Marine Biosciences. http://www.nrc-cnrc.gc.ca/eng/ibp/imb.html [cited]
98. Van De Riet, J et al (2011) Liquid chromatography post-column oxidation (PCOX) method for the determination of paralytic shellfish toxins in mussels, clams, oysters, and scallops: collaborative study. J AOAC International 94:1154–1176
99. Anderson DM (1997) Bloom dynamics of toxic *Alexandrium* species in the northeastern US. Limnol Oceanogr 42(5 Pt 2):1009–1022
100. Hurst JW (1975) History of paralytic shellfish poisoning on the Maine coast 1958–1974. In: First international conference on toxic dinoflagellate blooms. Massachusetts Science and Technology Foundation, Wakefield
101. Anderson DM, Glibert PM, Burkholder JM (2002) Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. Estuaries 25:704–726
102. Boesch DF et al (1997) Harmful algal blooms in coastal waters: options for prevention, control, and mitigation. NOAA Coastal Ocean Program Decision Analysis Series No.10. NOAA Coastal Ocean Office, Silver Spring, 46pp + appendix
103. Mulligan HF (1975) Oceanographic factors associated with New England red tide blooms. In: First international conference on toxic dinoflagellate blooms. Massachusetts Science and Technology Foundation, Wakefield
104. He R et al (2008) Historic 2005 toxic bloom of *Alexandrium fundyense* in the western Gulf of Maine: 2. coupled biophysical numerical modeling. J Geophys Res Oceans 113:C07040–C07042
105. McGillicuddy DJ Jr et al (2005) Mechanisms regulating large-scale seasonal fluctuations in *Alexandrium fundyense* populations in the Gulf of Maine: results from a physical-biological model. Deep Sea Res II 52:2698–2714
106. McGillicuddy DJ Jr et al (2003) A mechanism for offshore initiation of harmful algal blooms in the coastal Gulf of Maine. J Plankton Res 25(9):1131–1138
107. Anderson DM, Taylor CD, Armbrust EV (1987) The effects of darkness and anaerobiosis on dinoflagellate cyst germination. Limnol Oceanogr 32:340–351
108. Anderson DM et al (1982) Vertical and horizontal distributions of dinoflagellates cyst in sediments. Limnol Oceanogr 27:757–765
109. Sinclair GA et al (2006) Nitrate uptake by *Karenia brevis*. I. Influences of prior environmental exposure and biochemical state on diel uptake of nitrate. Mar Ecol Prog Ser 328:117–124
110. Sinclair GA et al (2006) Nitrate uptake by *Karenia brevis*. II. Behavior and uptake physiology in a nitrate-depleted mesocosm with a bottom nutrient source. Mar Ecol Prog Ser 328:125–131
111. Walsh JJ, Steidinger KA (2001) Saharan dust and Florida red tides: the cyanophyte connection. J Geophy Res Oceans 106(C6):11597–11612
112. Tester PA, Steidinger KA (1997) *Gymnodinium breve* red tide blooms: initiation, transport, and consequences of surface circulation. Limnol Oceanogr 42(5):1039–1051
113. Vargo GA et al (2008) Nutrient availability in support of *Karenia brevis* blooms on the central West Florida shelf: what keeps *Karenia* blooming? Cont Shelf Res 28(1):73–98
114. Baden DG, Mende TJ (1982) Toxicity of 2 toxins from the Florida red tide marine dinoflagellate *Ptychodiscus brevis*. Toxicon 20(2):457–461
115. Flewelling LJ et al (2005) Red tides and marine mammal mortalities. Nature 435(7043):755–756
116. Davis CC (1948) *Gymnodinium brevis* sp. nov a cause of discolored water and animal mortality in the Gulf of Mexico. Bot Gaz 109(3):358–360
117. Fire SE et al (2008) Prevalence of brevetoxins in prey fish of bottlenose dolphins in Sarasota Bay, Florida. Mar Ecol Prog Ser 368:283–294
118. Kirkpatrick B et al (2010) Inland transport of aerosolized Florida red tide toxins. Harmful Algae 9(2):186–189
119. Kirkpatrick B et al (2006) Environmental exposures to Florida red tides: effects on emergency room respiratory diagnoses admissions. Harmful Algae 5(5):526–533
120. Zaias J et al (2011) Repeated exposure to aerosolized brevetoxin-3 induces prolonged airway hyperresponsiveness and lung inflammation in sheep. Inhal Toxicol 23(4):205–211
121. Hetland RD, Campbell L (2007) Convergent blooms of *Karenia brevis* along the Texas coast. Geophys Res Lett 34(19):5
122. Flewelling LJ et al (2010) Brevetoxins in sharks and rays (Chondrichthyes, Elasmobranchii) from Florida coastal waters. Mar Biol 157(9):1937–1953
123. Gannon DP et al (2009) Effects of *Karenia brevis* harmful algal blooms on nearshore fish communities in southwest Florida. Mar Ecol Prog Ser 378:171–186
124. Kirkpatrick B et al (2004) Literature review of Florida red tide: implications for human health effects. Harmful Algae 3(2):99–115
125. Backer LC et al (2003) Recreational exposure to aerosolized brevetoxins during Florida red tide events. Harmful Algae 2(1):19–28
126. Fleming LE et al (2005) Initial evaluation of the effects of aerosolized Florida red tide toxins (Brevetoxins) in persons with asthma. Environ Health Perspect 113(5):650–657
127. Steidinger KA et al (1999) Harmful algal blooms in Florida. Unpublished technical report submitted to the Florida Harmful Algal Bloom Task Force. Florida Marine Research Institute, St. Petersburg, p 63
128. Hoagland P et al (2009) The costs of respiratory illnesses arising from Florida gulf coast *Karenia brevis* blooms. Environ Health Perspect 117(8):1239–1243
129. Morgan KL, Larkin SL, Adams CM (2010) Red tides and participation in marine-based activities: estimating the response of Southwest Florida residents. Harmful Algae 9(3):333–341
130. Morgan KL, Larkin SL, Adams CM (2009) Firm-level economic effects of HABS: a tool for business loss assessment. Harmful Algae 8(2):212–218

131. Plakas SM, Dickey RW (2010) Advances in monitoring and toxicity assessment of brevetoxins in molluscan shellfish. Toxicon 56(2):137–149
132. Stumpf R et al (2008) Hydrodynamic accumulation of *Karenia* on the west coast of Florida. Cont Shelf Res 28:189–213
133. Sinclair G (2008) Environmental, behavioral, and physiological constraints on the persistence of near-bottom populations of *Karenia brevis*. Department of marine, earth, and atmospheric sciences, North Carolina State University, Raleigh, p 181
134. Schofield O et al (1999) Optical monitoring and forecasting systems for harmful algal blooms: possibility or pipe dream? J Phycol 35(6):1477–1496
135. Janowitz GS, Kamykowski D (2006) Modeled *Karenia brevis* accumulation in the vicinity of a coastal nutrient front. Mar Ecol Prog Ser 314:49–59
136. Liu G, Janowitz G, Kamykowski D (2001) Influence of environmental nutrient conditions on *Gymnodinium breve* (Dinophyceae) population dynamics: a numerical study. Mar Ecol Prog Ser 213:13–37
137. Janowitz GS, Kamykowski D (1999) An expanded eularian model of phytoplankton environmental response. Ecol Model 118:237–247
138. Van Dolah FM, Leighfield TA (1999) Diel phasing of the cell-cycle in the Florida red tide dinoflagellate, *Gymnodinium breve*. J Phycol 35(6):1404–1411
139. Kamykowski D, Milligan EJ, Reed RE (1998) Biochemical relationships with the orientation of the autotrophic dinoflagellate *Gymnodinium breve* under nutrient replete conditions. Mar Ecol Prog Ser 167:105–117
140. Kamykowski D, Yamazaki H (1997) A study of metabolism-influence orientation in the diel vertical migration of marine dinoflagellates. Limnol Oceanogr 42(5):1189–1202
141. Dragovich A, Finucane J, Mays B (1961) Counts of red tide organisms, *Gymnodinium breve* and associated oceanographic data from Florida west coast, 1957–1959. In: U.S. Fish and Wildlife Service, Special. Science Report on Fish, Washington, DC, 175p
142. Steidinger KA et al (1998) Bloom dynamics and physiology of *Gymnodinium breve* with emphasis on the Gulf of Mexico. In: Anderson DM, Cembella AD, Hallegraeff GM (eds) Physiological ecology of harmful algal blooms. Springer, Berlin, pp 133–153
143. Masserini RT, Fanning KA (2000) A sensor package for the simultaneous determination of nanomolar concentrations of nitrite, nitrate, and ammonia in seawater by fluorescence detection. Mar Chem 68(4):323–333
144. Hecky RE, Kilham P (1988) Nutrient Limitation of phytoplankton in freshwater and marine environments: a review on the effects of enrichment. Limnol Oceanogr 33(4):796–822
145. Walsh J, Steidinger K (2001) Saharan dust and Florida red tides: the cyanophyte connection. J Geophys Res 106(C6):11,597–11,612
146. Mulholland MR et al (2006) Nitrogen fixation and release of fixed nitrogen by *Trichodesmium* spp. in the Gulf of Mexico. Limnol Oceanogr 51(4):1762–1776
147. Lester KM et al (2008) Zooplankton and *Karenia brevis* in the Gulf of Mexico. Cont Shelf Res 28(1):99–111
148. Sinclair GA, Kamykowski D (2008) Benthic-pelagic coupling in sediment-associated populations of *Karenia brevis*. J Plankton Res 30(7):829–838
149. Janowitz GS, Kamykowski D, Liu G (2008) A three-dimensional wind and behaviorally driven population dynamics model for *Karenia brevis*. Cont Shelf Res 28(1):177–188
150. Hasle GR (2002) Are most of the domoic acid-producing species of the diatom genus *Pseudo-nitzschia* cosmopolites? Harmful Algae 1:137–146
151. Smetacek V et al (2002) Mesoscale distribution of dominant diatom species relative to the hydrographical field along the Antarctic polar front. Deep-Sea Res II 49:3835–3848
152. Fryxell GA, Villac MC, Shapiro LP (1997) The occurrence of the toxic diatom genus *Pseudo-nitzschia* (Bacillariophyceae) on the West Coast of the USA, 1920–1996: a review. Phycologia 36:419–437
153. Qi Y, Wang J, Zheng L (1994) The taxonomy and bloom ecology of Pseudo-nitzschia on the coasts of China. In: IOC-WESTPAC third international scientific symposium, Bali, pp 88–95
154. Zou JZ, Zhou MJ, Zhang C (1993) Ecological features of toxic *Nitzschia pungens* Grunow in Chinese coastal waters. In: Smayda TJ, Shimizu Y (eds) Toxic phytoplankton blooms in the sea. Elsevier, Amsterdam
155. Evans KM et al (2004) Microsatellite marker development and genetic variation in the toxic marine diatom *Pseudo-nitzschia multiseries* (Bacillariophyceae). J Phycol 40(5):911–920
156. Evans KM, Hayes PK (2004) Microsatellite markers for the cosmopolitan marine diatom *Pseudo-nitzschia pungens*. Mol Ecol Notes 4(1):125–126
157. Evans KM, Kuhn SF, Hayes PK (2005) High levels of genetic diversity and low levels of genetic differentiation in North Sea *Pseudo-nitzschia pungens* (Bacillariophyceae) populations. J Phycol 41(3):506–514
158. Hubbard KA, Rocap G, Armbrust EV (2008) Inter- and intraspecific community structure within the diatom genus *Pseudo-nitzschia* (Bacillariophyceae). J Phycol 44:637–649
159. Orsini L et al (2004) Multiple rDNA ITS-types within the diatom *Pseudo-nitzschia delicatissima* (Bacillariophyceae) and their relative abundances across a spring bloom in the Gulf of Naples. Mar Ecol Prog Ser 271:87–98
160. Thessen AE, Bowers HA, Stoecker DK (2009) Intra- and inter-species differences in growth and toxicity of Pseudo-nitzschia while using different nitrogen sources. Harmful Algae 8:792–810
161. Bates SS (2000) Domoic-acid-producing diatoms: another genus added. J Phycol 36:978–985
162. Bates SS et al (1989) Pennate diatom *Nizschia pungens* as the primary source of domoic acid, a toxin in shellfish from eastern Prince Edward Island, Canada. Can J Fish Aquatic Sci 46:1203–1215
163. Lefebvre KA, Robertson A (2010) Domoic acid and human exposure risks: a review. Toxicon 56:218–230

164. Teitelbaum J, Carpenter S, Cashman NR (1990) Neurologic sequelae after ingestion of mussels contaminated with domoic acid. N Engl J Med 323:1632–1633
165. Peng YG et al (1994) Neuroexcitatory and the neurotoxic actions of the amnesic shellfish poison, domoic acid. Neuropharmacol Neurotoxicol 5:981–985
166. Gulland FMD et al (2002) Domoic acid toxicity in California sea lions (*Zalophus californianus*): clinical signs, treatment and survival. Vet Rec 150:475–480
167. Perl TM et al (1990) An outbreak of toxic encephalopathy caused by eating mussels contaminated with domoic acid. N Engl J Med 322:1775–1780
168. Todd ECD (1993) Domoic acid and amnesic shellfish poisoning: a review. J Food Prot 56:69–83
169. Wright JLC et al (1989) Identification of domoic acid, a neuroexcitatory amino acid, in toxic mussels from eastern Prince Edward Island. Can J Chem 67:481–490
170. Garrison DL et al (1992) Confirmation of domoic acid production by *Pseudo-nitzschia australis* (Bacillariophyceae) cultures. J Phycol 28:604–607
171. Lefebvre KA et al (2002) Domoic acid in planktivorous fish in relation to toxic *Pseudo-nitzschia* cell densities. Mar Biol 140:625–631
172. Scholin CA et al (2000) Mortality of sea lions along the central California coast linked to a toxic diatom bloom. Nature 403:80–84
173. Wekell JC et al (1994) Occurrence of domoic acid in Washington state razor clams (*Siliqua patula*) during 1991–1993. Nat Toxins 2:197–205
174. Fritz L et al (1992) An outbreak of domoic acid poisoning attributed to the pennate diatom *Pseudonitzschia australis*. J Phycol 28:439–442
175. Work TM et al (1993) Epidemiology of DA poisoning in brown pelicans (*Pelecanus occidentalis*) and Brandt's cormorants (*Phalacrocarax penicillatus*) in California. J Zoo Wildl Med 24:54–62
176. Sierra-Beltran AP et al (1997) Sea bird mortality at Cabo San Lucas, Mexico: evidence that toxic diatom blooms are spreading. Toxicon 35:447–453
177. Sierra-Beltran AP et al (1998) An overview of the marine food poisoning in Mexico. Toxicon 36:1493–1502
178. Lefebvre KA et al (1999) Detection of domoic acid in northern anchovies and California sea lions associated with an unusual mortality event. Nat Toxins 7(3):85–92
179. Goldstein T et al (2008) Novel symptomatology and changing epidemiology of domoic acid toxicosis in California sea lions (*Zalophus californianus*): an increasing risk to marine mammal health. Proc R Soc Biol Sci Ser B 275:267–276
180. Maucher JM, Ramsdell JS (2005) Domoic acid transfer to milk: evaluation of a potential route of neonatal exposure. Environ Health Perspect 115:1743–1746
181. Grattan LM et al (2007) Domoic acid neurotoxicity in native Americans in the pacific northwest: human health project methods and update. In: Fourth symposium on harmful algae in the US, Woods Hole
182. Trainer VL et al (2007) Recent domoic acid closures of shellfish harvest areas in Washington State inland waters. Harmful Algae 6:449–459
183. Trainer VL, Hickey BM, Horner RA (2002) Biological and physical dynamics of domoic acid production off the Washington coast. Limnol Oceanogr 47:1438–1446
184. Dyson K, Huppert DD (2010) Regional economic impacts of razor clam beach closures due to harmful algal blooms (HABs) on the Pacific coast of Washington. Harmful Algae 9:264–271
185. Bill BD et al (2006) The first closure of shellfish harvesting due to domoic acid in Puget Sound, Washington, USA. Afr J Mar Sci 28:437–442
186. Ayers D, Reed H (2004) Managing important recreational and commercial shellfish fisheries around harmful algal blooms. In: 2003 Georgia Basin/Puget sound research conference, Olympia
187. Hasle GR, Syvertsen EE (1997) *Marine Diatoms*. In: Tomas CR (ed) Identifying marine diatoms and dinoflagellates. Academic, San Diego
188. Hasle GR, Lundholm N (2005) *Pseudo-nitzschia seriata* f. *obtusa* (Bacillariophyceae) raised in rank based on morphological, phylogenetic and distributional data. Phycologia 44:608–619
189. Lundholm N et al (2003) A study of the *Pseudo-nitzschia pseudodelicatissima/cuspidata* complex (Bacillariophyceae): what is *P. pseudodelicatissima*? J Phycol 39:797–813
190. Lundholm N et al (2006) Inter- and intraspecific variation of the *Pseudo-nitzschia delicatissima* complex (Bacillariophyceae) illustrated by rRNA probes, morphological data and phylogenetic analyses. J Phycol 42(2):464–481
191. Fritz L (1992) The use of cellular probes in studying marine phytoplankton. Korean J Phycol 7:319–324
192. Rhodes LL (1998) Identification of potentially toxic *Pseudo-nitzschia* (Bacillariophyceae) in New Zealand coastal waters, using lectins. N Z J Mar Freshw Res 32:537–544
193. Rhodes LL et al (1998) Domoic acid producing *Pseudo-nitzschia* species educed by whole cell DNA probe-based and immunochemical assays. In: Reguera B et al (eds) Harmful algae. Xunta de Galicia and IOC of UNESCO, Santiago de Compostela, pp 274–277
194. Cho ES et al (1999) The rapid differentiation of toxic *Alexandrium* and *Pseudo-nitzschia* species using fluorescent lectin probes. J Korean Soc Oceanogr 35:167–171
195. Fraga S et al (1998) *Pseudo-nitzschia* species isolated from Galician waters: toxicity, DNA content and lectin binding assay. In: Harmful algae: eighth international conference on harmful algae, Vigo, pp 270–273
196. Rhodes L, Scholin C, Garthwaite I (1998) *Pseudo-nitzschia* in New Zealand and the role of DNA probes and immunoassays in refining marine biotoxin monitoring programmes. Nat Toxins 6:105–111
197. Andree KB et al (2011) Quantitative PCR coupled with melt curve analysis for detection of selected *Pseudo-nitzschia* spp. (Bacillariophyceae) from the northwestern Mediterranean Sea. Appl Environ Microbiol 77(5):1651–1659

198. Cho ES, Kodaki Y, Park JG (2001) The comparison of two strains of toxic *Pseudo-nitzschia multiseries* (Hasle) Hasle and non-toxic *Pseudo-nitzschia pungens* (Grunow) Hasle isolated from Chinhae Bay, Korea. Algae 16:275–285
199. Lundholm N et al (2002) Morphology, phylogeny and taxonomy of species within *Pseudo-nitzschia americana* complex (Bacillariophyceae) with descriptions of two new species, *Pseudo-nitzschia brasiliana* and *Pseudo-nitzschia linea*. Phycologia 41:480–497
200. Manhart JR et al (1995) *Pseudo-nitzschia pungens* and *P. multiseries* (Bacillariophyceae): nuclear ribosomal DNA's and species differences. J Phycol 31:421–427
201. McDonald SM, Sarno D, Zingone A (2007) Identifying *Pseudo-nitzschia* species in natural samples using genus-specific PCR primers and clone libraries. Harmful Algae 6(8):849–860
202. Miller P, Scholin CA (1996) Identification of cultured *Pseudo-nitzschia* (Bacillariophyceae) using species specific LSU rRNA-targeted probes. J Phycol 32:646–655
203. Miller P, Scholin CA (1998) Identification and enumeration of cultured and wild *Pseudo-nitzschia* (Bacillariophyceae) using species specific LSU rRNA-targeted fluorescent probes and filter-based whole cell hybridization. J Phycol 34:371–382
204. Scholin CA et al (1996) Identification of *Pseudo-nitzschia australis* (Bacillariophyceae) using rRNA-targeted probes in whole cell and sandwich hybridization formats. Phycologia 35:190–197
205. Scholin CA et al (1999) DNA probes and a receptor-binding assay for detection of *Pseudo-nitzschia* (Bacillariophyceae) species and domoic acid activity in cultured and natural samples. J Phycol 35:1356–1367
206. Scholin CA et al (1997) Detection and quantification of *Pseudo-nitzschia australis* in cultured and natural populations using LSU rRNA-targeted probes. Limnol Oceanogr 42:1265–1272
207. Scholin CA et al (1994) Ribosomal DNA sequences discriminate among toxic and non-toxic *Pseudo-nitzschia* species. Nat Toxins 2:152–165
208. Cho ES et al (2002) Monthly monitoring of domoic acid producer *Pseudo-nitzschia multiseries* (Hasle) Hasle using species-specific DNA probes and WGA lectins and abundance of *Pseudo-nitzschia* species (Bacillariophyceae) from Chinhae Bay, Korea. Bot Mar 45:364–372
209. Orsini L et al (2002) Toxic *Pseudo-nitzschia multistriata* (Bacillariophyceae) from the Gulf of Naples: morphology, toxin analysis and phylogenetic relationships with other *Pseudo-nitzschia* species. Eur J Phycol 37:247–257
210. Parsons ML et al (1999) *Pseudo-nitzschia* species (Bacillariophyceae) in Louisiana coastal waters: molecular probes field trials, genetic variability, and domoic acid analyses. J Phycol 35:1368–1378
211. Vrieling EG et al (1996) Identification of a domoic acid-producing *Pseudo-nitzschia* species (Bacillariophyceae) in the Dutch Wadden sea with electron microscopy and molecular probes. Eur J Phycol 31:333–340
212. Greenfield D et al (2008) Field applications of the second-generation environmental sample processor (ESP) for remote detection of harmful algae: 2006–2007. Limnol Oceanogr Method 6:667–679
213. Greenfield DI et al (2006) Application of the environmental sample processor (ESP) methodology for quantifying *Pseudo-nitzschia australis* using ribosomal RNA-targeted probes in sandwich and fluorescent in situ hybridization. Limnol Oceanogr Method 4:426–435
214. Scholin C et al (2009) Remote detection of marine microbes, small invertebrates, harmful algae and biotoxins using the environmental sample processor (ESP). Oceanography 22:158–167
215. AOAC (1990) Paralytic shellfish poison. Biological method. Final action. In: Hellrich K (ed) Official method of analysis. Association of Official Analytical Chemists (AOAC), Arlington, pp 881–882
216. Van Dolah FM et al (1997) A microplate receptor assay for the amnesic shellfish poisoning toxin, domoic acid, utilizing a cloned glutamate receptor. Anal Biochem 245:102–105
217. Garthwaite I et al (1998) An immunoassay for determination of domoic acid in shellfish and sea water. In: Reguera B et al (eds) Harmful algae. Xunta de Galicia and IOC of UNESCO, Santiago de Compostela, pp 559–562
218. Kawatsu K, Hamano Y, Noguchi T (1999) Production and characterization of a monoclonal antibody against domoic acid and its application to enzyme immunoassay. Toxicon 37:1579–1589
219. Smith DS, Kitts DD (1994) A competitive enzyme-linked immunoassay for domoic acid determination in human body fluids. Food Chem Toxicol 32(12):1147–1154
220. Smith DS, Kitts DD (1995) Enzyme Immunoassay for the determination of domoic acid in mussel extracts. J Agric Food Chem 43:367–371
221. Fernandez ML, Cembella AD (1995) Part B. Mammalian bioassays. In: Hallegraeff GM, Anderson DM, Cembella AD (eds) Manual on harmful marine microalgae. UNESCO, Paris, pp 213–228
222. Lawrence JF et al (1989) Liquid chromatographic determination of domoic acid in shellfish products using the paralytic shellfish extraction procedure of the association of official analytical chemists. J Chromatogr 462:349–356
223. Lawrence JF, Charbonneau CF, Menard C (1991) Liquid chromatographic determination of domoic acid in mussels, using AOAC paralytic shellfish poison extraction procedures: collaborative study. J Assoc Off Anal Chem 74(1):68–72
224. Ciminiello P et al (2005) Hydrophilic interaction liquid chromatography/mass spectrometry for determination of domoic acid in Adriatic shellfish. Rapid Commun Mass Spectrom 19(4):2030–2038
225. Hummert C, Reichelt M, Luckas B (1997) Automatic HPLC-UV determination of domoic acid in mussels and algae. Chromatographia 45:284–288
226. Quilliam MA et al (1989) High-performance liquid-chromatography of domoic acid, a marine neurotoxin, with application to shellfish and plankton. Int J Environ Anal Chem 36:139–154

227. European Union Reference Laboratory for Marine Biotoxins (ed) (2010) Standard operating procedure for determination of domoic acid (ADP toxins) in molluscs by UPLC-MS. European Union Reference Laboratory for Marine Biotoxins/ Agencia Espanola de Seguridad Alimentaria y Nutricion, Vigo
228. Rafuse C et al (2004) Rapid monitoring of toxic phytoplankton and zooplankton with a lateral-flow immunochromatographic assay for ASP and PSP toxins. In: Steidinger KA (ed) Harmful algae. Florida Fish and Wildlife Conservation Commission and Intergovernmental Oceanographic Commission of UNESCO, St. Petersburg
229. Turrell E et al (2008) Detection of *Pseudo-nitzschia* (Bacillariophyceae) species and amnesic shellfish toxins in Scottish coastal waters using oligonucleotide probes and the Jellet rapid test. Harmful Algae 7(4):443–458
230. Parsons ML, Dortch Q, Turner RE (2002) Sedimentological evidence of an increase in *Pseudo-nitzschia* (Bacillariophyceae) abundance in response to coastal eutrophication. Limnol Oceanogr 47(2):551–558
231. Dortch Q et al (2000) *Pseudo-nitzschia* spp. in the northern Gulf of Mexico: overview and response to increasing eutrophication. In: Symposium on harmful marine algae in the U.S. Marine Biological Laboratory, Woods Hole, p 27
232. McFadyen A, Hickey BM, Foreman MGG (2005) Transport of surface waters from the Juan de Fuca eddy region to the Washington coast. Cont Shelf Res 25:2008–2021
233. Bates SS et al (1991) Controls on domoic acid production by the diatom *Nitzschia pungens* f. *multiseries* in culture: nutrient and irradiance. Can J Fish Aquatic Sci 48:1136–1144
234. Pan Y, Subba Rao DV, Mann KH (1996) Changes in domoic acid production and cellular chemical composition of the toxigenic diatom *Pseudo-nitzschia multiseries* under phosphate limitation. J Phycol 32:371–381
235. Pan Y et al (1996) Effects of silicate limitation on production of domoic acid, a neurotoxin, by the diatom *Pseudo-nitzschia multiseries*. I. Batch culture studies. Mar Ecol Prog Ser 131:225–233
236. Pan Y et al (1996) Effects of silicate limitation on production of domoic acid, a neurotoxin, by the diatom *Pseudo-nitzschia multiseries*. II. Continuous culture studies. Mar Ecol Prog Ser 131:235–243
237. Van Apeldoorn ME, van Egmond HP, Speijers GJA (1999) Amnesic shellfish poisoning: a review. In: RIVM report 388802 019. National Institute of Public Health and the Environment, the Netherlands
238. Maldonado MT, Hughes MP, Rue EL (2002) The effects of Fe and Cu on growth and domoic acid production by *Pseudo-nitzschia multiseries* and *Pseudo-nitzschia australis*. Limnol Oceanogr 47:515–526
239. de Baar HJW et al (2005) Synthesis of iron fertilization experiments: from the iron age in the age of enlightenment. J Geophys Res Oceans 110:1–24
240. Marchetti A et al (2008) Identification and assessment of domoic acid production in oceanic *Pseudo-nitzschia* (Bacillariophyceae) from iron-limited waters in the northeast subartic Pacific. J Phycol 44(3):650–661
241. Bates SS et al (1995) Enhancement of domoic acid production by reintroducing bacteria to axenic cultures of the diatom *Pseudo-nitzschia multiseries*. Nat Toxins 3:429–435
242. Bates SS et al (2004) Interaction between bacteria and the domoic-acid-producing diatom *Pseudo-nitzschia multiseries* (Hasle) Hasle; can bacteria produce domoic acid autonomously? Harmful Algae 3:11–20
243. Kaczmarska I et al (2005) Diversity and distribution of epibiotic bacteria on *Pseudo-nitzschia multiseries* (Bacillariophyceae) in culture, and comparison with those on diatoms in native seawater. Harmful Algae 4:725–741
244. Brock TD (1973) Evolutionary and ecological aspects of the cyanophytes. In: Carr NG, Whitton BA (eds) The biology of blue-green algae. University of California Press, Berkeley, pp 487–500
245. Whitton BA, Potts M (eds) (2002) The ecology of cyanobacteria: their diversity in time and space. Academic, New York
246. Schofp JW (2002) The fossil record: tracing the roots of the cyanobacterial lineage. In: Whitton BA, Potts M (eds) The ecology of cyanobacteria: their diversity in time and space. Academic, New York, pp 13–35
247. Francis G (1878) Poisonous Australian lake. Nature 18:11–12
248. Codd GA, Bell SG, Brooks WP (1989) Cyanobacterial toxins in water. Water Sci Technol 21:1–13
249. Pilotto LS et al (1997) Health effects of exposure to cyanobacteria (blue-green algae) due to recreational water-related activities. Aust N Z J Public Health 21:562–566
250. Stewart I (2008) Cyanobacterial poisoning in livestock, wild animals, and birds–an overview. In: Hudnell HK (ed) Cyanobacterial harmful algal blooms: state of the science and research needs. Springer, New York, p 500
251. Teixera MGLC et al (1993) Gastroenteritis epidemic in the area of the Itaparica Dam, Bahia, Brazil. Bull Pan Am Health Organ 27:244–253
252. Tisdale ES (1931) Epidemic of intestinal disorders in Charleston, W.Va., occurring simultaneously with unprecedented water supply conditions. Am J Public Health 21:198–200
253. Turner PC et al (1990) Pneumonia associated with cyanobacteria. Br Med J 300:1440–1441
254. Metcalf JS, Codd GA (2004) Cyanobacterial toxins in the water environment. A review of current knowledge. Foundation for Water Research, Marlow
255. Chorus I, Bartram J (eds) (1999) Toxic cyanobacteria in water. A guide to their public health consequences, monitoring, and management. E & FN Spon, London (on behalf of the World Health Organization)
256. Falconer IR (1993) Measurement of toxins from blue-green algae in water and foodstuffs. In: Falconer IR (ed) Algal toxins in seafood and drinking water. Academic, New York, pp 165–176
257. Cheng YS et al (2007) Characterization of aerosols containing microcystin. Mar Drugs 5:136–150

258. Falconer IR, Buckley TH (1989) Tumour promotion by *Microcystis* sp., a blue-green algae occurring in water supplies. Med J Australia 150:351–352
259. Falconer IR, Humpage AR (1996) Tumour promotion by cyanobacterial toxins. Phycologia 35(Suppl 6):74–79
260. Fujiki H, Sueoka E, Suganuma M (1996) Carcinogenesis of microcystins. In: Watanabe MF et al (eds) Toxic microcystis. CRC Press, Boca Raton, pp 203–232
261. Humpage AR et al (2000) Microcystins (cyanobacterial toxins) in drinking water enhance the growth of aberrant crypt foci in the mouse colon. J Toxicol Environ Health A 61:155–165
262. Zhou L, Yu H, Chen K (2002) Relationship between microcystin in drinking water and colorectal cancer. Biomed Environ Sci 15(2):166–171
263. World Health Organization (2008) Guidelines for drinking-water quality: incorporating the 1st and 2nd addenda. World Health Organization, Geneva
264. Burns J (2008) Toxic cyanobacteria in Florida waters. In: Hudnell HK (ed) Cyanobacterial harmful algal blooms: state of the science and research needs. Springer, New York, p 500
265. Westrick JA (2008) Cyanobacterial toxin removal in drinking water treatment processes and recreational water. In: Hudnell HK (ed) Cyanobacterial harmful algal blooms: state of the science and research needs. Springer, New York, p 500
266. Carmichael WW (2001) Assessment of blue-green algal toxins in raw and finished drinking water. AWWA Research Foundation and American Water Works Association, Denver
267. Chorus I, Fastner F (2001) Recreational exposure to cyanotoxins. In: Chorus I (ed) Cyanotoxins, occurrence, causes, consequences. Springer, Heidelberg, pp 190–199
268. Miller MA et al (2010) Evidence for a novel marine harmful algal bloom: cyanotoxin (microcystin) transfer from land to sea otters. PLoS One 5(9):e12576
269. Wiegand C, Pflugmacher S (2005) Ecotoxicological effects of selected cyanobacterial secondary metabolites a short review. Toxicol Appl Pharmacol 203:201–218
270. Funari E, Testai E (2007) Human health risk assessment related to cyanotoxins exposure. Crit Rev Toxicol 38(2):97–125
271. Henricksen P et al (1997) Detection of an anatoxin-a(s)-like anticholinesterase in natural blooms and cultures of cyanobacteria/blue-green algae from Danish lakes and in the stomach contents of poisoned birds. Toxicon 35(1111):901–913
272. Matsunaga S et al (1989) Anatoxin-a(s), a potent anticholinesterase from *Anabaena flos-aquae*. J Am Chem Soc 111(494):8021–8023
273. Mahmood NA, Carmichael WW (1986) Paralytic shellfish poison produced by the freshwater cyanobacterium *Aphanizomeno flos-aquae* NH-5. Toxicon 24:175–186
274. Stewart I, Schluter PJ, Shaw GR (2006) Cyanobacterial lipopolysaccharides and human health – a review. Environ Health A Global Access Sci Source 5:7
275. Smith JL, Boyer GL, Zimba PV (2008) A review of cyanobacterial odorous and bioactive metabolites: impacts and management alternatives in aquaculture. Aquaculture 280:5–20
276. Fujiki H et al (1985) A blue-green alga from Okinawa contains aplysiatoxins, the third class of tumour promoters. Jpn J Cancer Res 76:257–259
277. Shimizu Y (1996) Microalgal metabolites: a new perspective. Annu Rev Microbiol 50:431–465
278. Fujiki H et al (1990) New tumour promoters from marine natural products. In: Hall S, Strichartz G (eds) Marine toxins. Origin, structure and molecular pharmacology. American Chemical Society, Washington, DC, pp 232–240
279. Gorham PR, Carmichael WW (1988) Hazards of freshwater blue-green algae (cyanobacteria). In: Lembi CA, Waaland JR (eds) Algae and human affairs. Cambridge University Press, New York, pp 404–431
280. Aguirre AA et al (2006) Hazards associated with the consumption of sea turtle meat and eggs: a review for health care workers and the general public. EcoHealth 3(3):141–153
281. Yasumoto Y (1998) Fish poisoning due to toxins of microalgal origins in the Pacific. Toxicon 36:1515–1518
282. Papapetropoulos S (2007) Is there a role for naturally occurring cyanobacterial toxins in neurodegeneration? The beta-*N*-methylamino-*L*-alanine (BMAA) paradigm. Neurochem Int 50:998–1003
283. Wilson JM et al (2002) Behavioral and neurological correlates of ALS-parkinsonian dementia complex in adult mice fed washed cycad flour. Neuromol Med 1:207–221
284. Rao SD et al (2006) BMAA selectively injures motor neurons via AMP/kainate receptor activation. Exp Neurol 201:244–252
285. Lobner D et al (2007) Beta-*N*-methyl-amino-*L*-alanine enhances neurotoxicity through multiple mechanisms. Neurobiol Dis 25(2):360–365
286. Buenz EJ, Howe CL (2007) Beta-methylamino-alanine (BMAA) injures hippocampal neurons in vivo. Neurotoxicology 28(3):702–704
287. Liu XQ et al (2010) Selective death of cholinergic neurons induced by beta-methylamino-L-alanine. Neuroreport 21(1):55–58
288. Karlsson O et al (2009) Selective brain uptake and behavioral effects of the cyanobacterial toxin BMAA (β-*N*-methylamino-*L*-alanine) following neonatal administration to rodents. Toxicol Sci 109(2):286–295
289. Wilde SB et al (2005) Avian vacuolar myelinopathy linked to exotic aquatic plants and a novel cyanobacterial species. Environ Toxicol 20(3):348–353
290. Bidigare RR et al (2009) Cyanobacteria and BMAA: possible linkage with avian vacuolar myelinopathy (AVM) in the southeastern United States. Amyotroph Lateral Scler 10(Suppl 2):71–73
291. Fischer JR, Lewis-Weis LA, Tate CM (2003) Experimental vacuolar myelinopathy in red-tailed hawks. J Wildl Dis 39(2):400–406
292. Birrenkot AH et al (2004) Establishing a food-chain link between aquatic plant material and avian vacuolar myelinopathy in mallards (*Anas plytyrhynchos*). J Wildl Dis 40(3):485–494

293. Murch SJ, Cox PA, Banack SA (2004) A mechanism for slow release of biomagnified cyanobacterial neurotoxins and neurodegenerative disease in Guam. Proc Natl Acad Sci U S A 101:12228–12231
294. Esterhuizen M, Downing TG (2008) Beta-*N*-methylamino-*L*-alanine (BMAA) in novel South African cyanobacterial isolates. Ecotoxicol Environ Saf 71:309–313
295. Johnson HE et al (2008) Cyanobacteria (*Nostoc commune*) used as a dietary item in the Peruvian highlands produces the neurotoxic amino acid BMAA. J Ethnopharmacol 118:159–165
296. Li A et al (2010) Detection of the neurotoxin BMAA within cyanobacteria isolated from freshwater in China. Toxicon 55:947–953
297. Metcalf JS et al (2008) Co-occurrence of beta-*N*-methylamino-*L*-alanine, a neurotoxic amino acid with other cyanobacterial toxins in British waterbodies, 1990–2004. Environ Microbiol 10(3):702–708
298. Cox PA et al (2005) Diverse taxa of cyanobacteria produce beta-*N*-methylamino-*L*-alanine, a neurotoxic amino acid. Proc Natl Acad Sci U S A 102:5074–5078
299. Caller TA et al (2009) A cluster of amyotrophic lateral sclerosis in New Hampshire: a possible role for toxic cyanobacteria blooms. Amyotroph Lateral Scler 10(Suppl 2):101–108
300. Tucker CS (2000) Off-flavor problems in aquaculture. Rev Fish Sci 8(1):45–88
301. Dodds WK et al (2009) Eutrophication of U.S. freshwaters: analysis of potential economic damages. Environ Sci Technol 43:12–19
302. World Health Organization (2003) Guidelines for safe recreational water environments. Volume 1: coastal and fresh waters. World Health Organization, Geneva
303. Steffensen DA (2008) Economic cost of cyanobacterial blooms. In: Hudnell HK (ed) Cyanobacterial harmful algal blooms: state of the science and research needs, vol 619, Advances in experimental medicine and biology. Springer, New York, 500pp
304. Kenefick SL et al (1992) Odorous substances and cyanobacterial toxins in praire drinking water sources. Water Sci Technol 25:147–154
305. Paerl HW (2008) Nutrient and other environmental controls of harmful cyanobacterial blooms along the freshwater-marine continuum. In: Hudnell HK (ed) Cyanobacterial harmful algal blooms: state of the science and research needs, vol 619, Advances in experimental medicine and biology. Springer, New York, 500pp
306. Garcia-Villada L et al (2004) Occurrence of copper resistant mutants in the toxic cyanobacteria *Microcystis aeruginosa*: characterization and future implications in the use of copper sulphate as algaecide. Water Res 38:2207–2213
307. Hitzfeld BC, Hoeger SJ, Dietrich DR (2000) Cyanobacterial Toxins: removal during drinking water treatment, and human risk assessment. Environ Health Perspect 108(6):113–122
308. Hoeger SJ, Hitzfeld BC, Dietrich DR (2005) Occurrence and elimination of cyanobacterial toxins in drinking water treatment plants. Toxicol Appl Pharmacol 203:231–242
309. Hoeger SJ et al (2004) Occurrence and elimination of cyanobacterial toxins in two Australian drinking water treatment plants. Toxicon 43(14248):639–649
310. Jones G, Gurney S, Rocan D (1996) Water quality/toxic algae study interim report: summary of the 1995 field season results. Manitoba Environment, 4 June 1996
311. Lambert TW, Holmes CFB, Hrudey SE (1996) Adsorption of microcystin-LR by activating carbon and removal in full-scale water treatment. Water Res 30(6):1411–1422
312. Wannemacher JRW et al (1993) Treatment for removal of biotoxins from drinking water. US Army Biomedical Research and Development Laboratory, Fort Detrick, Frederick
313. James H, Lloyd A (2002) Blue-green algae and their toxins – Great Britain's perspective. In: Proceedings of the health effects of exposure to cyanobacteria toxins: state of the science, Sarasota
314. Kim S-C, Lee D-K (2005) Inactivation of algal blooms in eutrophic water of drinking water supplies with the photocatalysis of TiO2 thin film on hollow glass beads. Water Sci Technol 52(9):145–152
315. Ueno Y et al (1996) Detection of microcystins, a blue-green algal hepatotoxin, in drinking water sampled in Haimen and Fusui, endemic areas of primary liver cancer in China, by highly sensitive immunoassay. Carcinogenesis 17:1317
316. Koskenniemi K et al (2007) Quantitative real-time PCR detection of toxic *Nodularia* cyanobacteria in the Baltic Sea. Appl Environ Microbiol 73(7):2173–2179
317. Rinta-Kanto JM et al (2005) Quantification of toxic *Microcystis* spp. during the 2003 and 2004 blooms in western Lake Erie using quantitative real-time PCR. Environ Sci Technol 39(11):4198–4205
318. Codd GA et al (2001) Analysis of cyanobacterial toxins by physiochemical and biochemical methods. J AOAC Int 84:1625–1635
319. Dorr FA et al (2010) Methods for detection of anatoxin-a(s) by liquid chromatography coupled to electrospray ionization-tandem mass spectrometry. Toxicon 55:92–99
320. Fischer WJ et al (2001) Congener-independent immunoassay for microcystins and nodularins. Environ Sci Technol 35:4849–4856
321. Metcalf JS, Bell SG, Codd GA (2001) Colorimetric immunoprotein phophatase inhibition assay for specific detection of microcystins and nodularins of cyanobacteria. Appl Environ Microbiol 67:904–909
322. Anderson WB, Slawson RM, Mayfield CI (2002) A review of drinking-water-associated endotoxin, including potential routes of human exposure. Can J Microbiol 48:567–587
323. Carmichael WW (1992) Occurrence of toxic cyanobacteria. US EPA. Agency, Cincinnati, OH, pp 15–26
324. Oberholster PJ, Botha A-M, Grobbelaar JU (2004) Microcystis aeruginosa: source of toxic microcystins in drinking water. African J Biotech 3:159–168
325. Carmichael WW (2008) A world overview one-hundred, twenty-seven years of research on toxic cyanobacteria – where do we go from here? In: Hudnell HK (ed)

Cyanobacterial harmful algal blooms: state of the science and research needs, vol 619, Advances in experimental medicine and biology. Springer, New York, 500 pp

326. Burns J, Williams C, Chapman A (2002) Cyanobacteria and their toxins in Florida surface waters. In: Proceedings of the health effects of exposure to cyanobacteria toxins: state of the science, Sarasota
327. Giddings M et al (2002) Cyanobacterial toxins: the development and evaluation of method to determine microcystin levels in Canadian water supplies. In: Proceedings of the health effects of exposure to cyanobacteria toxins: state of the science, Sarasota
328. Boyer GL (2008) Cyanobacterial toxins in New York and the lower great lakes ecosystems. In: Hudnell HK (ed) Cyanobacterial harmful algal blooms: state of the science and research needs. Springer, New York, p 500
329. Fristachi A et al (2007) A preliminary exposure assessment of microcystins from consumption of drinking water in the United States. Lake Reserv Manag 23(2):203–210
330. Osswald J et al (2007) Toxicology and detection methods of the alkaloid neurotoxin produced by cyanobacteria, anatoxin-a. Environ Int 33:1070–1089
331. Falconer IR (2005) Cyanobacterial toxins of drinking water supplies: cylindrospermopsins and microcystins. CRC Press, Florida
332. Walsby AE, Schanz F, Schmid M (2006) The Burgundy-blood phenomenon: a model of buoyancy change explains autumnal waterblooms by *Planktothrix rubescens* in Lake Zurich. New Phytol 169:109–122
333. Metting B, Pyne JW (1986) Biologically-active compounds from microalgae. Enzyme Microb Technol 8:386–394
334. Sellner KG (1997) Physiology, ecology, and toxic sproperties of marine cyanobacteria blooms. Limnol Oceanogr 42(5 Pt 2): 1089–1104

Books and Reviews

Anderson DM (2009) Approaches to monitoring, control and management of harmful algal blooms (HABs). Ocean Coast Manag 52(7):342–347

Babin M et al (2005) New approaches and technologies for observing harmful algal blooms. Oceanography 18(2):210–227

Bauer M et al (2010) The importance of human dimensions research in managing harmful algal blooms. Front Ecol Environ 8(2):75–83

Fleming LE et al (2011) Review of Florida red tide and human health effects. Harmful Algae 10(2):224–233

Graneli E, Turner JT (eds) (2006) Ecology of harmful algae. Springer, Berlin/Heidelberg

Hoagland P et al (2002) The economic effects of harmful algal blooms in the United States: estimates, assessment issues, and information needs. Estuaries Coast 25(4):819–837

Hudnell HK (ed) (2008) Advances in experimental medicine and biology, vol 619. Springer, New York

Hudnell HK (2010) Harmful algal blooms and natural toxins in fresh and marine waters – exposure, occurrence, detection, toxicity, control, management and policy. Toxicon 55(5):1024–1034

Hudnell HK et al (2010) Freshwater harmful algal bloom (FHAB) suppression with solar powered circulation (SPC). Harmful Algae 9:208–217

Landsberg JH (2002) The effects of harmful algal blooms on aquatic organisms. Rev Fish Sci 10(2):113–390

Lewis RJ, Poli M (2010) Toxins in seafood. Toxicon 56(2):107–258 (Special Issue)

Secher S (2009) Measures to control harmful algal blooms. The Plymouth Student Scientist 2(1):212–227

Shumway S, Rodrick G (eds) (2009) Shellfish safety and quality. Woodhead, Cambridge

Valiela I et al (1997) Macroalgal blooms in shallow estuaries: controls and ecophysiological and ecosystem consequences. Limnol Oceanogr 42(5 Pt 2):1105–1118

Hazardous Materials Characterization and Assessment

JULIE M. SCHOENUNG, CARL W. LAM
Department of Chemical Engineering and
Materials Science, University of California,
Davis, Davis, CA, USA

Article Outline

Glossary
Definition of Subject
Introduction
Hazardous Materials Characterization
Hazardous Materials Assessment Metrics
Future Directions
Bibliography

Glossary

Hazardous materials assessment Systematic review or aggregation methods to consider evidence for material toxicity and hazards to support waste management or pollution prevention priority setting for high-impact substances.

Hazardous materials characterization Identification of environmental, health, and physical hazard traits in materials.

Hazard traits Environmental, health, and physical adverse effects posed by materials to surroundings and living organisms.

Material life cycle The journey a material goes through during its entire life including stages of raw material production, product manufacturing, use, and end-of-life.

Materials selection A design process by which materials are compared and chosen for use according to their functional and environmental attributes.

Definition of the Subject

Some materials can be safely produced with minimal environmental and human health concerns during its use or disposal. Other materials are hazardous to manufacture or use and, when disposed, can contaminate and persist in the environment. These toxicity issues can depend on the chemical traits of the substances, such as polymers, metals and other compounds, in question. Characterization and assessment methods are needed to correctly identify substances of concern and to evaluate in a systematic way the degree of hazard they pose to ecological systems and human health. Building upon toxicological studies, hazardous materials management requires assessment tools that integrate toxicity data to support decision making for the proper handling, treatment, and/or elimination of toxic substances from industrial processes and manufactured products.

Historically, through regulatory or corporate efforts, environmental protection has addressed hazardous materials management primarily as an end-of-pipe treatment step to mitigate emissions and their impacts. Hazardous materials governance with this overarching precedence can potentially distribute the adverse environmental effects across all stages of the products' life cycle including raw material production, manufacturing, use, and end-of-life. As a progressive approach, emerging paradigms of "green engineering" and "green chemistry" seek to eliminate or substitute hazardous materials with inherently safer alternatives during the design of products and processes. Therefore, with the support of hazardous materials characterization and assessment tools, material options can be rigorously compared with consideration of hazard traits during materials selection in the design phase.

Introduction

The number of substances humans can use and synthesize from the natural world is nearly limitless. Ranging from pesticides, catalysts, building materials to electronics, these chemicals and materials help to create the many technological novelties that sustain infrastructure and activities in the modern world. Yet, over the last few decades, despite the clear benefits these new substances offer to improve products and processes in society, there is an increasing awareness of the hazards they pose to humans and natural ecosystems during their life cycles [1, 2]. Current converging trends in international environmental regulations attest to this growing global interest to control and mitigate the toxicity impacts caused by the use of hazardous materials in commercial industries [3].

To address these issues, hazardous materials characterization and assessment frameworks provide a critical first step to identify and compare the multitude of hazard traits that are associated with substances. The communication of material hazard information is essential to highlight and prioritize waste streams for pollution prevention during waste management or remediation and also necessary for achieving safe and sustainable products and processes by design [4]. Clearly, the unwanted dissipation of hazardous materials from use and production increases the chance of pollution and wastes accumulating in environmental media, which can be irreversible. Ecological systems and human populations are threatened in different regions where these materials are mismanaged, making the long-term consequences of unrestrained toxic substances use severe.

An important example of hazardous materials misuse is provided in the context of metallic lead (Pb) in commercial markets. Lead has been one of the most widely used substances worldwide, with applications as a pure metal, an alloy, or as an additive to organic substances, ceramics or glass materials [5]. The human health toxicity of lead has long been recognized and neurological effects can be witnessed even with very low levels in blood, with children being particularly susceptible [6, 7]. Historically, two applications turned lead toxicity into a widespread problem. The first was the use of lead as a brightness enhancer for paints in homes, and second was the use of organo-lead (tetraethyl lead) as a motor fuel antiknock agent [8]. Both have shown to contribute lead into soil and dust contamination and to play a significant role in human exposure [9]. In the U.S., it was not until the late 1970s when the phase-out of lead additives eventually demonstrated a marked decline of lead blood levels in the population [10]. From a preventative standpoint, this is a clear situation of where appropriate procedures for hazardous materials characterization and assessment at the onset would have promulgated lead's toxicity issues and motivated for reduction in lead exposure and unintended toxicity impact to populations.

At present, there is not a unified consensus on best practices to benchmark and compare materials and chemicals to determine whether a material is "greener" or "safer" than another in terms of its hazards. Fundamentally, the main reason for this nonuniformity is because of the many ways by which substances can be characterized based on their hazard traits. Depending on the substance in question, different hazard criteria may be a focus for consideration.

The aim of this article is to illustrate the multifaceted definition of hazard in materials and how various toxicity data can be aggregated within hazardous materials assessments. The first portion of this article defines what properties make materials hazardous to begin with. Hazard issues from the life cycle perspective of materials are also described. This discussion is followed up with particular examples for plastics and metals. The second portion of this article seeks to provide an overview of currently available assessment metrics and methodologies utilized to compare materials considering the defined hazard traits. Such tools can support decision making during materials selection and priority setting for hazardous materials management.

Hazardous Materials Characterization

Hazard Traits in Materials

Hazardous materials characterization (or identification) is the determination of what type of endpoint adverse effects will be caused by an exposure or release of a substance. Before the discussion on hazardous materials assessment tools, it is necessary to first establish what these hazard traits of materials are.

One of the most prominent hazard trait classification systems is the Globally Harmonized System of Classification and Labeling of Chemicals (GHS) [11]. GHS represents an internationally motivated system with the aim of unifying hazard classification and labeling standards in different nations and regions in the world. Within GHS, classification of hazard criteria falls into three areas:

1. Physical hazards
2. Human health hazards
3. Environmental (or ecosystem) hazards

Figure 1 shows a compilation of the chemical hazard traits for these GHS hazard categories.

Physical hazards are classifications that define the danger posed by a material based on its physical state (gas, liquid, or solid). These hazards can imply many or overlapping endpoints of concern for a substance. These include considerations such as flammability, corrosion, and oxidizing properties of substances and other hazard attributes of a substance that can cause bodily harm or damage to surroundings. An explosive substance is capable of producing gas by chemical reaction at such a temperature, pressure, and speed to cause damage to surroundings. Flammable substances are readily combustible, or may cause or contribute to fire under certain atmospheric conditions. Oxidizing substances refer to materials that if provided oxygen can cause or contribute to the combustion of itself and other materials. Self-reactive substances are thermally unstable substances with the potential to undergo a strongly exothermic reaction even without air. A pyrophoric substance has potential to ignite through contact with air. Self-heating substances have the potential to self-heat by reaction with air and without an energy source. A substance that will materially damage metals is corrosive to metal. More detailed definitions for specific traits in the physical hazard categories can be found in reference [11].

Health hazards, of central importance for hazard classifications, are a series of adverse health effects posed to humans when they are exposed to a toxic substance. Key exposure pathways are inhalation, oral, or dermal routes. These can derive from epidemiological or animal studies. Acute toxicity implies the intrinsic property of a substance to cause adverse effect in humans after short-term exposure. Skin corrosion or irritation refers to substances that can respectively cause irreversible or reversible damage in skin. Likewise, substances reflecting serious eye damage or irritation can cause irreversible or reversible damage in the eye. Respiratory or skin sensitization refers to a substance that induces hypersensitivity of the airways

Physical Hazard

- Explosives
- Flammable gases
- Flammable aerosols
- Oxidizing gases
- Gases under pressure
- Flammable liquids
- Flammable solids
- Self-reactive substances
- Pyrophoric liquids
- Pyrophoric solids
- Self-heating substances
- Substances which, in contact with water, emit flammable gases
- Oxidizing liquids
- Oxidizing solids
- Organic peroxides
- Corrosive to metals

Health Hazard

- Acute toxicity
- Skin corrosion/irritation
- Serious eye damage/eye irritation
- Respiratory or skin sensitization
- Germ cell mutagenicity
- Carcinogenicity
- Reproductive toxicology
- Target organ systemic toxicity – single exposure
- Target organ systemic toxicity – repeated exposure
- Aspiration toxicity

Environmental hazard

- Hazardous to the aquatic environment
  - Acute aquatic toxicity
  - Chronic aquatic toxicity
    - Bioaccumulation potential
    - Rapid degradability

Hazardous Materials Characterization and Assessment. Figure 1
Physical, health, and environmental hazard classifications within GHS [11]

following inhalation or an allergic response following skin contact, respectively. Germ cell mutagenicity classifies substances liable to cause mutations in organisms. Reproductive toxicity implies a substance's potential to adversely affect sexual function and fertility in adults, as well as healthy development in children. Aspiration toxicity refers to substances that can cause acute adverse effects following inhalation. Repeated or single exposure target organ/systemic toxicity refer to substances that can trigger organ specific toxicity effects. Carcinogenicity implies a substance's capability to induce cancer or increase its incidence in humans.

Generally in GHS, each of these health classifications is assigned a "category" number to indicate its damage severity. For example in the acute toxicity category, substances are assigned category 1 (most severe) to 5 (low acute toxicity) classification on the basis of lethal dose or concentration 50% (LD50) (oral, dermal) or LC50 (inhalation) from animal studies. These indicators are defined as the dose or concentration that will be lethal for 50% of tested animals (which are typically rodents). Carcinogens are classified differently according to the categorization scheme provided in Table 1.

Hazardous Materials Characterization and Assessment. Table 1 GHS carcinogenicity categorization scheme [11]

GHS carcinogenicity category	Definition
1A	Known human carcinogen (based on human evidence)
1B	Presumed human carcinogen (based on demonstrated animal carcinogenicity)
2	Suspected carcinogen (limited evidence of human or animal carcinogenicity)

For the environmental hazards category, substances are classified by acute (short term) and chronic (long term) aquatic toxicity data. The basis for acute aquatic toxicity is the median dose of a substance that will likely cause irreversible damage to species (e.g., fish, crustacean, or other aquatic organisms). This is represented by the LC50 or effective concentration for 50% (EC50) in aquatic species. Chronic aquatic toxicity indicators (e.g., chronic EC50s) imply the potential of a substance to cause adverse effect with continuous exposures throughout the lifecycle of the organism at a specified concentration. Note that for chronic toxicity of substances with lack of rapid degradability, other chemical parameters would be required for classification such as the logarithm of the substance's octanol–water partition coefficient (K_{ow}), to indicate how easily a chemical may be distributed in water bodies [12], and the Bioconcentration Factor (BCF), to characterize the tendency for a chemical to accumulate in organisms [13]. Categorization schemes are also available for environmental hazards in GHS according to Table 2.

Beyond GHS, there are also other international agencies and organizations that provide other hazard trait classifications, particularly in the area of human health hazards. For this, substances are usually described based on their maximum exposure limits and carcinogenic potential to humans. Notable organizations that develop human exposure limit classifications include: the U.S. Occupational Safety and Health Administration (OSHA) who provide Permissible Exposure Limits (PELs) to regulate chemical concentrations in the workplace [14]; the American Conference of Governmental Industrial Hygienists (ACGIH) who develop similar workplace Threshold Limit Values (TLVs) [15]; the U.S. National Institute for Occupational Safety and Health (NIOSH) who provides Recommended Exposure Limits (RELs) which are recommended air concentration limits to protect employees in their working lifetimes [16]; and the German Research Foundation who presents their own maximum occupational exposure limits (MAKs) [17]. Other organizations that identify carcinogenic potential in substances include the International Agency for Research on Cancer (IARC) [18], the European Union (EU) Council [19], and the U.S. National Toxicology Program (NTP) [20].

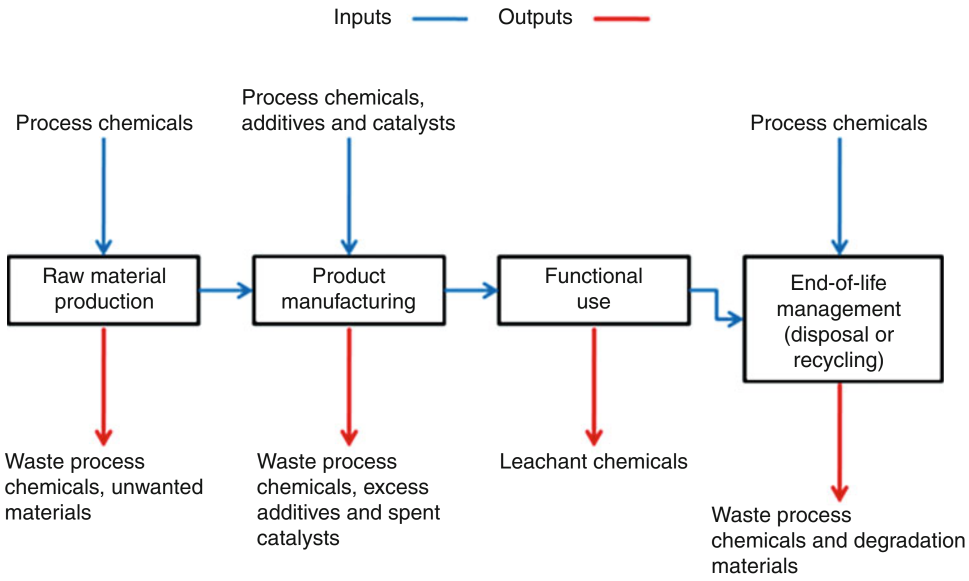
Historically, there have been limited international organizations that provide threshold concentrations or classifications for ecological toxicity in the natural environment. However, there are a few available that are focused on aquatic impacts. From America, the U.S. Environmental Protection Agency (USEPA) provides values for Total Maximum Daily Load (TMDL) as limits on the amount of a pollutant that a body of water in the natural environment can receive daily while still meeting national water quality standards [21], and the State of California lists maximum freshwater concentrations for inland surface waters and enclosed bays and estuaries [22]. From the EU, the German water pollution classification rating (WGK) represents a substance's potential to be hazardous in water bodies [23].

Life Cycle Contributions to Material Hazards

For some substances, the substance itself does not present inherent hazards. These substances may become of concern, however, when, at some point in the material's life cycle, secondary materials react with them or when their fate and persistence in surrounding environments become problematic. Another issue that cannot be disregarded in hazard characterization is that materials are often manufactured by adding or combining other constituents (e.g., metal alloys), meaning they can exhibit multiple hazard traits deriving from the diverse elemental or molecular groups present. Individually, these substances embody a separate range of physical, health, and environmental hazards. A truly complete

Hazardous Materials Characterization and Assessment. Table 2 Acute and chronic GHS environmental hazards categorization scheme [11]

Acute category I	Acute category II	Acute category III	
EC50 $\leq$ 1.00 mg/l	EC50 $>$ 1.00 but $\leq$ 10.0 mg/l	EC50 $>$ 10.0 but $<$ 100 mg/l	
Chronic category I	Chronic category II	Chronic category III	Chronic category IV
EC50 $\leq$ 1.00 mg/l and log K_{ow}[b] $\geq$ 4 unless BCF[c] $<$ 500	EC50 $>$ 1.00 but $\leq$ 10.0 mg/l and log K_{ow} $\geq$ 4 unless BCF $<$ 500 and unless chronic toxicity $>$ 1 mg/l	EC50 $>$ 10.0 but $\leq$ 100.0 mg/l and log K_{ow} $\geq$ 4 unless BCF $<$ 500 and unless chronic toxicity $>$ 1 mg/l	EC50 $>$ 100 mg/l and log K_{ow} $\geq$ 4 unless BCF $<$ 500 and unless chronic toxicity $>$ 1 mg/l


Inputs
Outputs
Process chemicals
Process chemicals, additives and catalysts
Process chemicals
Raw material production
Product manufacturing
Functional use
End-of-life management (disposal or recycling)
Waste process chemicals, unwanted materials
Waste process chemicals, excess additives and spent catalysts
Leachant chemicals
Waste process chemicals and degradation materials

Hazardous Materials Characterization and Assessment. Figure 2
Material life cycle hazard profile from associated inputs and outputs

understanding of hazard traits implied by a material's use requires factoring in these hazard contributions throughout the material life cycle, starting from raw material production to product manufacturing, to use and, ultimately, in end-of-life management.

Figure 2 depicts the potential contributions to the life cycle hazard profile of a material. Different input substances are introduced in separate stages, leading to outputs of potentially hazardous substances. From this viewpoint, the choice of using a particular material requires understanding of the hazards attributed to input and output substances throughout a material's life. Inputs include processing chemicals that are added and utilized to develop a material product. This begins with raw material production and occurs throughout product manufacturing. Some of these additions may not be innocuous and can lead to serious health or environmental concerns if inappropriately managed. Material production of gold, for example, commonly requires the use of cyanide (an acutely toxic chemical) for ore extraction processes [24]. End-of-life materials management can also require using additional chemicals during hazardous waste treatment processes. A common strategy for wastewater treatment is the introduction of hydroxides and sulfides to remove heavy metal contaminants from concentrated solutions through precipitation [25]. Outputs are substances emitted in the material life cycle including waste process chemicals or spent chemicals (e.g., catalysts). Chemicals may leach from the material itself during use or product disposal. Outputs can also be represented as hazardous waste streams emitted during end-of-life disposal, recycling, or other waste management processes.

Ultimately, what this schematic highlights is that, in addition to consideration of the material itself, primary and intermediate chemicals used during manufacturing and, also importantly, chemicals generated at end-of-life, are dynamic contributors to embodied material hazards. Polymers can potentially degrade into their monomer form, which may be more toxic (e.g., polyvinyl chloride to vinyl chloride) [26]. In addition, toxic chemicals can be generated by reaction during waste management processing. For example, natural decomposition of organic materials in landfills can create air pollution such as methane, carbon dioxide, nitrogen, sulfur, and volatile organic compounds [27]. Also, a side effect of waste-to-energy incineration processes for materials is that this can lead to the generation of unwanted toxic substances in fly ash such as heavy metals and other combustion by-products [28]. Although recycling will help to return materials into the industrial cycle, toxic substances may also be generated through industrial recovery processes.

In order to characterize the hazard attributes of a material, in parallel to what is actually in the material's composition, the related substances that appear throughout the material life cycle also factor into consideration. To provide some examples of these issues in real substances, a general overview of hazard traits is applied to characterize polymers, specifically plastics, and metals in the following section.

Case Study Examples: Hazard Characterization of Plastics and Metals

Plastics Plastics materials, a specific type of polymer utilized ubiquitously in global industries, are macromolecules comprised of covalently bonded repeating subunits (or monomers). Table 3 presents the hazard classifications in alignment with GHS categories [29] for various monomer types including acrylonitrile-butadiene-styrene (ABS), polycarbonate, polyethylene, polyethylene terephthalate, polypropylene, polystyrene, and polyvinyl chloride. This table shows a range of hazard endpoints that can be of concern to environmental and human health as well as physical hazard categories. Of the plastics considered, ABS, polystyrene, and polyvinyl chloride have monomers that are indicated to have carcinogenic potential in GHS categories. ABS, polycarbonate, polyethylene terephthalate, polystyrene, and polyvinyl chloride monomers are classified to be toxic to reproductive systems. In terms of environmental hazard, with the exception of polypropylene due to lack of GHS classification, all of the plastics are categorized to have acute toxicity for aquatic organisms. In addition, physical hazards, as expected for the hydrocarbon-based materials, suggest issues with flammability for some monomers.

The multiple hazard traits posed by monomers used in plastics shown in Table 3 comprise only a segment of the implied life cycle hazards. As suggested by the previous description on hazardous materials characterization, understanding the full impact that plastic materials present requires also understanding the additional input and output substances that occur throughout a material's life cycle.

During production, the majority of plastics are developed from the reaction of petroleum-derived chemicals such as ethylene, benzene, propylene, and butadiene [30]. Additional processing converts primary petrochemicals into intermediate chemicals such as ethylbenzene formed from ethylene and benzene. Then, the monomers are created from the intermediate chemicals such as styrene from ethylbenzene. The plastic (i.e., polystyrene) can then be produced by polymerization reaction of the monomer. In this process, catalysts are often introduced to increase the rate of polymerization.

Some instances of plastic manufacturing require the blending of two independent polymers together (such as polycarbonate and ABS) in order to synthesize a new material (i.e., a copolymer) with different material performance characteristics [31]. Also, chemical additives are often incorporated to modify some mechanical or physical property of the material. These additives may be in the form of flame retardants, ultraviolet and thermal stabilizers, and other processing or functional enhancers [32]. It is not unusual for one type of plastic to contain multiple additives. ABS, for example, can simultaneously utilize ethoxylated fatty alkylamines as an antistatic agent, sulfonyl semicarbazides or azodicarbonamides as foaming aids, calcium stearates as internal lubricants, and fatty amides as a mold release aid, all of which would be in addition to traditional halogenated compounds as flame retardants [32]. The incorporation of all of these substances within the plastic material directly increases the complexity when performing hazard characterization of such plastics [33].

The leaching of residual additives from materials into the environment is a major issue for humans and the ecological systems during use and disposal. Because of this, environmental legislation has emerged to limit toxins (i.e., brominated flame retardants and lead (Pb)) used in commercial materials, which affects the manufacturing of plastics [34]. Of note, however, is that some additives that exhibit toxicity traits have yet to be regulated. Reproductive and developmental toxins such as bisphenol A from polycarbonate bottles [35, 36] and phthalates from polyvinyl chloride plastics are some widely publicized examples [37, 38].

During the end-of-life processing of plastics, in addition to issues posed by monomers and additives that may leach out from these materials, other toxic substances can also be generated due to inappropriate

Hazardous Materials Characterization and Assessment. Table 3 Monomers of common plastic types and their hazard traits in alignment with GHS [29]

Plastic type (monomers marked in parentheses)	Hazard traits of monomers		
Acrylonitrile-butadiene-styrene (acrylonitrile, butadiene, styrene)	Acrylonitrile	Butadiene	Styrene
	Human hazard: Category 2 carcinogenicity Category 2 toxic to reproduction	Human hazard: Category 2 carcinogenicity Category 2 toxic to reproduction	Human hazard: Category 2 carcinogenicity Category 1 toxic to reproduction
	Environmental hazard: Category 3 acute aquatic toxicity	Environmental hazard: Category 3 acute aquatic toxicity	Environmental hazard: Category 1 acute aquatic toxicity
	Physical hazard: Category 2 flammable liquid	Physical hazard: Category 2 flammable liquid	Physical hazard: Category 3 flammable liquid
Polycarbonate (bisphenol A)	Bisphenol A		
	Human hazard: Category 2 toxic to reproduction		
	Environmental hazard: Category 1 acute aquatic toxicity		
	Physical hazard: Not applicable		
Polyethylene (ethylene)	Ethylene		
	Human hazard: Category 3 systemic toxicity (narcotic effects)		
	Environmental hazard: Category 3 acute aquatic toxicity		
	Physical hazard: Category 1 flammable gas		
Polyethylene terephthalate (ethylene terephthalate)	Ethylene	Terephthalate	
	Human hazard: Category 3 systemic toxicity (narcotic effects)	Human hazard: Category 4 acute oral toxicity Category 2 toxic to reproduction	
	Environmental hazard: Category 3 acute aquatic toxicity	Environmental hazard: Category 3 acute aquatic toxicity	
	Physical hazard: Category 1 flammable gas	Physical hazard: Not applicable	
Polypropylene (propylene)	Propylene		
	GHS classification not available		

Hazardous Materials Characterization and Assessment. Table 3 (Continued)

Plastic type (monomers marked in parentheses)	Hazard traits of monomers		
Polystyrene (styrene)	Styrene		
	Human hazard: Category 2 carcinogenicity Category 1 toxic to reproduction		
	Environmental hazard: Category 1 acute aquatic toxicity		
	Physical hazard: Category 3 flammable liquid		
Polyvinyl chloride (vinyl chloride)	Vinyl chloride		
	Human hazard: Category 1 carcinogenicity Category 2 toxic to reproduction		
	Environmental hazard: Category 1 acute aquatic toxicity		
	Physical hazard: Category 1 flammable gas		

hazardous materials management procedures. These substances may also have significant implications on the environmental damage potential of plastic goods. A well-documented case is the generation of dioxins and furans occurring during waste incineration of polyvinyl chloride electronic cables to recover copper [39]. These particular by-product chemicals have been established to be persistent, bioaccumulative, and carcinogenic to exposed populations [40].

From this overall perspective, it is clear that plastics represent a class of materials for which many variations in inputs and outputs during the material life cycle can affect the comprehensive hazard profile. A compilation of hazard traits for the wide range of substances of concern involved during an entire plastic material life cycle has never been generated, but is necessary to comprehensively gauge impacts posed by additives, process chemicals, and degradation behavior of plastics.

Metals A similar high-level discussion on hazard characterization can be presented for metals. From a toxicological standpoint, metals themselves represent an interesting scenario for hazard characterization because at the right concentrations, some metals are essential to living organisms and ecological systems [41]. However, chronic low-dose exposures of others can lead to severe health and environmental effects. Table 4 shows a compilation of metals with an indication of whether they are or are not essential to humans.

In addition, different metals valence states (chemical species) can change the hazard attributes of the metal. There are specific chemical processes that can transform metals from one valence state to another, oftentimes enabling the formation of inorganic and organic substances with different toxicity properties. For instance, chromium can exist in a trivalent state (III), which is essential for proper development in mammals, or a hexavalent state (VI), which is highly toxic to organisms [42, 43]. In general, hazardous metals require special attention because, unlike organic chemicals or polymer materials, they can neither be created nor destroyed by biological or chemical reactions. This may lead to a greater chance for accumulation in various environmental compartments and also in organisms [44].

The metals themselves represent a range of environmental, health, and physical hazard traits. In the human health hazard category, at least six metals (arsenic, cadmium, chromium (VI), beryllium, nickel, and lead) are accepted or suspected to be human carcinogens in one form or another for particular routes of exposure, according to GHS characterization [29]. For certain metals (e.g., lead and mercury), other human hazard traits such as toxicity to reproduction and development have been demonstrated [29]. With regard to the environmental hazard characterization of metals listed in Table 4, copper, zinc, barium, beryllium, cadmium, and thallium have been classified to have toxicity potential with chronic exposures to aquatic organisms (GHS category 3 or 4) [29]. There is a lack of ecological toxicity data to characterize the ecological hazard traits of other metal species. For the physical hazard category, the majority of metals are not classified in GHS due to the nature of solid metals.

From a life cycle point of view, production of primary metals creates hazardous waste output issues, especially for localities where mining and extraction of metals from ores occur [45]. Chemical pollution (and therefore hazards attributed to outputs) can arise from various routes during mineral processing. A typical large-scale production can leave high metal concentrations such as copper, zinc, lead, cadmium, and arsenic in rivers, groundwater, and sediments [45]. Slag, a necessary by-product of ore-smelting processes, is comprised of metal oxides, elemental metals, and various sulfides that can contaminate soils. Many metal ores are naturally formed as sulfides (e.g., nickel, copper, and lead), which, if exposed to oxygen and water during mining operations, oxidize to create sulfuric acid [46]. In addition, chemicals are often added to ores during mineral processing and these can pollute ecological systems if not properly managed. Examples include the use of concentrated acids for the leaching of copper oxides and other metal oxides [46] or, as mentioned, cyanide for gold. As a whole, these production materials and process chemicals are contributors to the upstream impact of metals.

Hazardous Materials Characterization and Assessment. Human essential and nonessential metals and their valence states [44]

Essential elements	Nonessential metals	
Cobalt II, III	Aluminum III	Nickel II
Chromium III	Antimony III, V	Niobium V
Copper 0, I, II	Arsenic III, V	Palladium 0, II
Iron II, III	Barium II	Platinum 0, II, IV
Magnesium II	Beryllium II	Rubidium I
Manganese II, IV	Bismuth III, V	Silicon IV
Molybdenum IV, VI	Boron III	Silver 0, I, II
Selenium II, IV, VI	Cadmium II	Strontium II
Zinc II	Cesium I	Tellurium II, IV, VI
	Chromium VI	Thallium I, III
	Gallium III	Tin II, IV
	Germanium IV	Titanium IV
	Gold 0, I, III	Tungsten VI
	Indium III	Uranium IV, VI
	Lead II, IV	Vanadium III, V
	Lithium I	Zirconium IV
	Mercury 0, I, II	

During manufacturing of metal materials, different metals can be combined to create alloys with modified performance properties. This can be analogous to the case of plastic blends and additives to enhance material functionality. Chromium, for example, is often introduced into various grades of aluminum alloys to improve corrosion resistance [47]. Like plastic additives, these "solute" metals may sometimes be supplemented only in minor quantities to the pure metal to achieve the desired property enhancement. As a whole, these additional metals contribute not only toward manufacturing hazards but also in end-of-life when disposal or recycling operations must manage these different metals simultaneously.

What is seen here is that even with consideration of just the materials themselves, such as plastic monomers and elemental metals, diverse hazard traits can arise under the environmental, human, and physical hazard categories. Factoring in additional considerations for plastics and metals, such as their additives, alloys, and compounds used for production and manufacturing from a life cycle perspective, makes the task of comprehensive hazard characterization of complex materials very challenging. Future efforts in the arena of "green engineering" and "green chemistry" need to include the development of hazardous materials assessment methods to identify and aggregate nonuniform hazards and toxicity information so that decision makers can benchmark materials and substances in a systematic way, while accounting for their entire life cycle.

Hazardous Materials Assessment Metrics

A consequence of technology evolution is the creation of increasingly complex waste streams and materials emitted from the industrial society into the natural environment. To this end, hazardous materials assessment metrics are necessary for the comparative evaluation of the hazard traits represented by the many new materials and substances being developed in assorted industry sectors such as energy products, semiconductors, automotive, chemicals, electronics, pharmaceuticals, plastics, and countless others [48]. This is a demanding task because decision making for the proper industrial management of materials can be clouded by the assorted and numerous hazard traits of substances that need to be considered, as discussed above.

The hazardous material assessment tools available today seek to address this issue by providing approaches that consider the available data on hazards so that substances of primary concern can be identified in a transparent and methodical way. Industrial waste stream priority setting is a clear application for these methods since they can help guide environmental decisions for pollution prevention [49–51]. Simultaneously, as a promising new paradigm in the chemical and materials engineering community, these tools can help support the safer design of processes and products by identifying high-impact materials before they are incorporated into a product [52]. Material choices in product design can greatly influence end-of-life waste management as well as implications for hazardous waste generated during manufacturing. Hazardous materials assessment metrics can enable a designer or engineer to evaluate inherent material hazards to humans and ecosystems in lieu of other material properties during product or process design.

Key approaches in this area have been developed to consider hazard information of materials with the objective of standardizing screening of high-impact substances and to streamline comparative evaluation. These tools provide hazard benchmarking for individual agents, but often require hazard and toxicity data collection (characterization) to be performed during the assessment. A summary of established tools, including a comparison of their structure and application, is provided in Table 5. Further details on each method are provided below. Note that because the foundation hazard and toxicity data may vary from one tool to another, clarification of relevant toxicological terminology is provided as needed.

Life Cycle Impact Assessment Tools

Life cycle impact assessment (LCIA) is used in life cycle analysis (LCA) to characterize the impacts of material flows (releases) of processes or products [53]. LCIA involves development of a multimedia and multipathway characterization model to capture the fate, exposure, and potency of a substance in order to determine its contribution to a given impact category (such as potential for ecological or human health toxicity). Commonly, results from these models are reported as "characterization factors" (CFs), which represent a material's impact in a specific category per unit of release (or emission). LCIA is generally conducted by multiplying the amount of release (or emission) by the CFs to quantify the magnitude of impact attributed to the substance. Therefore, impact category contributions from a wide variety of substances can be effectively summed and compared under a standard basis. Two prominent LCIA methods are summarized here.

Tool for the Reduction and Assessment of Chemical and Other Environmental Impacts (TRACI) Version 1.0 Although the Tool for the Reduction of Chemical and Other Environmental Impacts (TRACI) (version 1.0) is originally designed for use in LCIA [54], it has also recently seen wider application as a general pollution prevention and design for environment indicator toolset to quantify environmental impact potential of substances [49, 50, 55–59]. TRACI CFs for particular substances can be used to quantify the major environmental impact effects such as global warming, ozone depletion, acidification, eutrophication, smog formation, ecological toxicity (ecotoxicity), and human health–related effects. For this discussion, only human health and ecotoxicity effects are discussed due to an interest on quantifying toxicity hazards of substances.

TRACI human health CFs are represented by Human Toxicity Potentials (HTPs) [60]. HTPs are

Hazardous Materials Characterization and Assessment. Table 5 Comparison of different aspects of hazardous materials assessment tools

Method		TRACI	USETox^{TM}	RSEI	CHEMS-1	TPI	Green Screen
Application		Life cycle impact assessment	Life cycle impact assessment	Population hazard and risk screening	Hazard-based chemical prioritization	Comparative materials selection	Chemicals screening
Complexity for practitioners to apply		High	High	Low	Medium-high	Medium	Medium-high
Quantitative		Yes	Yes	Yes	Yes	Yes	No
		Toxicity data					
Hazard categories	Human health	Yes; human RfC/RfDs for non-carcinogens and cancer potency factors for carcinogens	Yes; ED50 for carcinogens, ED50 extrapolated from NOEL for non-carcinogens	Yes; human RfC/RfDs for non-carcinogens and cancer potency factors for carcinogens	Yes; rodent LD50/LC50, IARC and USEPA carcinogenicity classification	Yes; German MAKs, EU carcinogenic classification	Yes; various
	Environment	Yes; concentration-to-source ratio from CalTOX and impact-to-concentration ratio from PNEC	Yes; aquatic organisms EC50	No	Yes; fish LC50, fish NOEL, rodent LD50	Yes; German water hazard classification	Yes; aquatic organisms LC50/EC50/IC50; GHS categories for chronic and acute aquatic toxicity
	Physical or other safety issues	No	No	No	No	Yes; Risk phrases	Yes; GHS categories for flammability and explosives
	Fate and Exposure	Yes; CalTox for multimedia multi-pathway modeling based on U.S. parameters	Yes; Matrix algebra framework for multimedia multi-pathway modeling	Yes; AERMOD model of air releases	Yes; with indicators to account for	No	Yes; with indicators to account for
				First-order decay model for water releases	(a) Bioaccumulation: BCF		(a) Bioaccumulation: BCF, log k_{ow}
				RSEI makes inherent assumptions on probability of exposure pathways in human populations based on U.S. parameters	(b) Persistence: BOD half-life in water, chemical half-life in water		(b) Persistence: chemical half-life in sediments and water

Hazardous Materials Characterization and Assessment. Table 5 (Continued)

Method		TRACI	USETox^{TM}	RSEI	CHEMS-1	TPI	Green Screen
Method to convert to hazard value		Normalized with respect to chemicals such as benzene or toluene (not bounded)	Not normalized (not bounded)	Algorithms to convert chemical toxicity data to toxicity weights (not bounded)	Linear function (0–5 for environmental and health categories; 1–2.5 for fate/exposure categories)	Algorithm to convert to numeric scale (0–100)	Integer scales (1–4)
Method to aggregate hazard categories		Does not aggregate hazard categories	Does not aggregate hazard categories	Uses most sensitive pathway for adverse effect to represent toxicity weight	User choice of weighting factors to aggregate total hazard value	Equivalence between health, environmental, and physical	Benchmark scheme based on: very high, high, moderate, or low classification of hazards
Output of method		Characterization Factors	Characterization Factors	Toxicity weights (hazard-based) Risk scores (risk-based)	Hazard values	TPI scores	Benchmark tiers

RfD reference dose, *RfC* reference concentration, *ED50* effect dose 50%, *NOEL* no observable effect level, *LD50* lethal dose 50%, *LC50* lethal concentration 50%, *MAK* maximum allowed concentrations for workers, *PNEC* predicted no-effects concentration, *EC50* effective concentration 50%, *GHS* Globally Harmonized System, *IC50* inhibitory concentration 50%, *BCF* bioconcentration factor, *BOD* biological oxygen demand, $\log k_{ow}$ log octanol–water coefficient

cancer and non-cancer (all other adverse health effects) toxicity potentials associated with a chemical's release into select media (i.e., air or water) relative to that associated with the release of a reference chemical. Calculations for HTPs utilize the CalTOX model (Multimedia Total Exposure Model for Hazardous Waste Sites) to determine the pollutant's fate and transport characteristics as a function of the average acceptable daily dose or cancer potency factor for substances [61]. CFs for ecotoxicity potentials are determined by combining two components: (1) a concentration-to-source ratio (CSR) from the CalTOX model and (2) an impact-to-concentration ratio (ICR) from the predicted no-effects concentration (PNEC) (based on fraction of species adversely affected), to estimate potential terrestrial and aquatic toxicity impact relative to a reference chemical's media-specific release [54]. Reference chemicals used to indicate toxicological equivalency in TRACI include benzene for cancer potential, toluene for non-cancer potential, and 2,4-dichlorophenoxyacetic acid (2,4 D) for ecotoxicity potential.

A key distinction of TRACI is that it utilizes input parameters within CalTOX representing average conditions in the U.S., making its application as an international LCIA tool limited. A new methodological development of TRACI (version 2.0) has been recently updated to recommend adopting global characterization factors from another tool known as USETox^{TM} as described below [62].

USETox^{TM}: The UNEP-SETAC Toxicity Model The USETox^{TM} method is another LCIA method focused on toxicity hazard characterization. It is unique because its development derives from scientific consensus among international toxicity experts from the United Nations Environment Program (UNEP) and Society for Environmental Toxicology and Chemistry (SETAC) [63]. The objective of this tool is to harmonize the existing LCIA tools and resolve discrepancies between prominent impact analysis models. At its foundation, USETox^{TM} utilizes a matrix algebra approach for multimedia and pathway modeling, reflecting fate, exposure, and effect potency of substances in order to calculate CFs (see Table 5). The tool provides global CFs for substances released to urban air, nonurban air, freshwater, seawater, natural soil, and agricultural soil, for use in LCIA studies.

USETox^{TM} human health toxicity CFs are reported as a change in adverse effect probability due to change in lifetime intake of a pollutant from released quantities (cases/kg). These are also called Comparative Toxicity Units (CTU_h/kg), to highlight the comparative nature of the CFs. Ecological toxicity CFs reflect the change in the Potentially Affected Fraction (PAF) of species due to a change in concentration from released quantities (PAF^*m^3/kg or CTU_e/kg). As a further development, USETox^{TM} marks some CFs as "interim" to describe uncertainty in values for substances [64]. The details of the parameters considered for human health and environmental categories are provided in Table 5.

Integrated Screening Tools

The following tools are "screening" approaches developed by separate entities and organizations to simplify the process of aggregating hazard information for chemical assessments. These are different than LCIA tools in that their objective is not to quantify actual impacts of releases but to provide a screening perspective of environmental impact.

Risk-Screening Environmental Indicators (RSEI) Although not a formal risk assessment, USEPA's Risk-Screening Environmental Indicators (RSEI) model utilizes a system in which the toxicity of a given chemical can be quantified by a dimensionless "risk score" to provide a screening-level perspective for comparing chemical emissions' health impacts on human populations [65, 66]. RSEI focuses on releases and is tailored toward evaluation of U.S. Toxics Release Inventory (TRI) data, a national compilation of annual emissions information from manufacturing industries. RSEI is similar to LCIA-type tools in that it incorporates the amount of chemical released, chemical toxicity, fate and transport through the environment, exposure route, and population effects during evaluation (Table 5). The model uses the most potent chronic health endpoint in considering human health impacts due to exposure. "Toxicity weights," calculated from acceptable daily dose values and cancer potency factors, are developed as a chemical-specific toxicity valuation system within RSEI to determine risk scores. As compared to other tools, an important drawback within RSEI is that ecological toxicity is not explicitly addressed.

Toxic Potential Indicator (TPI) by Fraunhofer IZM Based on European-based regulated toxicity and hazards information, the Fraunhofer IZM Toxic Potential Indicator (TPI) method was originally designed to support materials selection for products and processes [67–71].

The computational structure of the TPI method requires three main input variables to reflect environmental, health, and other physical safety considerations for materials:

- Risk phrases (R-phrases)
- Occupational exposure limits based on maximum workplace concentration (MAK) or European Union (EU) carcinogenic classification
- Water hazard classification (WGK)

Risk phrases are defined in the Annex III of European Union (EU) Commission Directive 67/548/EEC: Nature of Special Risks Attributed to Dangerous Substances and Preparations, specifying the multiple hazards and safety issues presented by a material ranging from flammability potential to ecosystem impact potential [72]. Human health hazards are captured with MAK values (maximum concentration of a material allowed in the workplace) and European Union carcinogenic classifications [17, 19]. The German water pollution classification value, WGK, represents impacts posed to the environmental hazard category [23]. Exposure and fate of substances are not captured in TPI.

Because the tool is based on EU regulations, the methodology has also seen limited application on an international basis. However, there has been some ongoing effort to update the TPI method to other nations with case examples of Taiwan and Japan [70, 73].

Chemical Hazard Evaluation for Management Strategies (CHEMS-1) A chemical ranking and scoring method entitled "Chemical Hazard Evaluation for Management Strategies" (CHEMS-1) has been developed as a screening tool to provide a relative comparison of chemical hazards to human health and the environment [74–76]. The purpose of CHEMS-1 is to place chemical release data into perspective by evaluating both the toxic effects of chemicals and the potential exposure to those chemicals. Initially, CHEMS-1 was developed to identify high-impact chemicals within TRI. In principle however, this method can also be extended to support materials decision making in any product and process development.

CHEMS-1 captures human health effects with the use of chronic and acute toxicity data based on animal studies such as LD50s and also carcinogenic classification published by IARC. Environmental hazard effects are derived from rat LD50s for terrestrial toxicity and fish LC50s for aquatic toxicity (Table 5). Exposure potential is captured by persistence and bioaccumulation indicators such as the Bioconcentration Factor (BCF) and hydrolysis half-life. These indicators for various hazard categories are used as input parameters in linear functions to quantify "hazard values" that represent the impacts in the human health, environmental, and exposure categories. The summation of these impacts provides a "total hazard value" for a substance. It is also possible to integrate user weighting preferences for the different impact contributions to calculate for total hazard values.

Green Screen for Safer Chemicals Green Screen for Safer Chemicals is a chemical evaluation tool created by the Clean Production Action [77, 78]. This tool deviates from the others in that it does not try to enumerate hazard scores from hazard traits. The Green Screen approach is very comprehensive in the sense that a wide range of hazard traits is considered including human health, ecological toxicity, and fate as well as physical properties of substances. All of these have been identified through expert judgment and reviews of existing hazard classification schemes, including the GHS, the US EPA's Design for Environment program [79], and other classifications from different regional origins [77]. The resulting criteria include many international qualitative and quantitative thresholds. Within the Green Screen approach, a rubric is provided to help classify the hazard trait information by assigning a level of concern (low, medium, high, very high) for different hazard categories. The tool then considers these levels of concern with additional decision logic to generate an overall substance score in terms of integer "benchmarks." Scores can range from Benchmark 1, meaning a chemical of high concern, to Benchmark 4, a "safer" and preferred substance. Of the aforementioned screening tools presented, only Green Screen captures the degradation of materials, requiring assessments also on those degradation substances.

Case Study Examples: Hazard Assessment for Nickel and Bisphenol A

To provide a sense of the how the outputs of these assessment approaches can differ, nickel and bisphenol A are chosen to exemplify output scores of these tools for a metal and plastic monomer, respectively.

In terms of the LCIA-based tools, Table 6 gives the CF values that derive from TRACI version 1.0 and USETox$^{\text{TM}}$ for cancer, non-cancer, and ecotoxicity effects categories due to air emissions. It is not expected that the CFs be comparable between methods and this is confirmed as orders of magnitude differences are observed for the reported values for these substances in the select effect categories shown. In addition to using dissimilar units to represent CFs, these variations can be attributed to differences in methodologies, such as models used to account for fate, exposure, and effect considerations.

Despite these incompatibilities, CFs between methods show a similar trend when comparing the two substances within their respective set. Generally, nickel is expected to rank with a higher impact for all effect categories because it has higher burden to human health and to ecological systems. TRACI and USETox$^{\text{TM}}$ human health and ecotoxicity CFs for the substances confirms this. Note that because bisphenol A is a non-carcinogen, its cancer effects CFs are essentially zero for both TRACI and USETox$^{\text{TM}}$ methods. CFs for other media releases (such as water and soil) can also be used to quantify impacts attributed to these substances.

In parallel, indicators for nickel and bisphenol A are also provided based on results from the integrated screening assessment tools: RSEI, CHEMS-1, TPI, and Green Screen (Table 7). It is useful to keep in mind that, as described in previous summaries, there are inherent differences by which these screening tools account for hazard traits of substances. These fundamental variabilities are mainly derived from using dissimilar toxicity information, or other safety factors to represent a material's impact during hazard assessment. In addition, each screening tool has its own implicit weighting mechanism to aggregate hazard trait information.

RSEI toxicity weights for these substances, which do not account for fate or exposure considerations and only consider human health effects, show orders of magnitude variation. This is explained by the fact that RSEI uses the most sensitive health endpoint to represent hazard. Nickel is anticipated to be carcinogenic in humans, making its toxicity weight significantly greater in RSEI when compared to bisphenol A. Further use of these toxicity weights in RSEI enables determination of risk scores if population exposure is modeled [65].

CHEMS-1 hazard value results are aggregated into total hazard values and are normalized from 0 to 100

Hazardous Materials Characterization and Assessment. Table 6 TRACI and USETox$^{\text{TM}}$ characterization factors for nickel and bisphenol A [54, 63]

	Air – cancer effects		Air – non-cancer effects		Air – ecotoxicity effects	
Chemical	TRACI (kg benzene-equivalents)	USETox$^{\text{TM}}$ (CTU_h/kg)	TRACI (kg toulene-equivalents)	USETox$^{\text{TM}}$ (CTU_h/kg)	TRACI (kg 2,4 D - equivalents)	USETox$^{\text{TM}}$ (CTU_e/kg)
Nickel	3.61	5.18×10^{-5}	8.26×10^{3}	2.91×10^{-6}	1.30×10^{2}	6.08×10^{3}
Bisphenol A	0	0	4.21	2.53×10^{-7}	0	9.37×10^{1}

Hazardous Materials Characterization and Assessment. Table 7 Integrated screening assessment tool hazard indicators for nickel and bisphenol A [65, 74, 77]

Chemical	RSEI (toxicity weights)	CHEMS-1 (normalized total hazard value)	TPI (TPI score)	Green screen
Nickel	930,000	59.0	38.0*	Benchmark 1*
Bisphenol A	20	6.5*	8.7*	Benchmark 1

*Author-determined scores are marked with an asterisk with respect to the highest attainable value for each hazard category considered [74]. This can most directly compare against the TPI method, which by default presents substance scores with the same range from 0 to 100 [67]. Similar to other indicators, both CHEMS-1 and TPI show again nickel having a higher aggregated human health and environmental impact.

The Green Screen ratings for these substances present an interesting scenario where these two substances are assigned an equivalent Benchmark 1 (highest burden) score. These Green Screen results are caused by separate mechanisms. In the case of nickel, Benchmark 1 is the outcome of the decision logic because it is an anticipated carcinogen. For bisphenol A, it is because the substance is classified as an endocrine disruptor [80]. In Green Screen, these hazard traits effectively shift a substance into the highest level of concern for the human health category according to the method's rubric. Determination of substance scores using the other screening methods (i.e., RSEI, CHEMS-1, and TPI) does not account for endocrine disruption.

Caveats on the Tools

As highlighted in Table 5, different tools have been developed to consider select hazard traits of substances so that substances can be compared on the basis of a defined standard. However, there are key differences when looking at these tools in terms of their application. LCIA tools are designed for application within LCA studies, as they attempt to capture actual effects based on a media release through multimedia and pathway modeling of different chemicals. With the exception of RSEI, the screening tools do not attempt to model fate and exposure behavior. Because LCA-type studies entail the analysis of life cycle substance flows for products, such assessments are generally more complex to conduct. Screening tools are designed to compare individual materials. Depending on the number of hazard traits considered in various environmental, health, and physical hazard categories, they are generally easier to apply. Green Screen is an exception to this since assessments require significant consideration of many toxicity data sources due to its comprehensiveness.

Also, these tools inherently aggregate hazard scores differently. For example, of the methods described, only CHEMS-1 provides a user preference to weigh hazard categories. The LCIA tools essentially provide CFs without a mechanism to weigh contributions from health or ecological toxicity. RSEI simply uses the most potent endpoint for substances in the health category to calculate toxicity weights. Green Screen has its own internal logic system based on decision rules to value hazard traits. The TPI method by default weighs environmental, health, and physical hazards equivalently within its computational structure.

In the end, it is important for designers to recognize these differences and intricacies when applying these tools to identify materials of concern.

Future Directions

There is a clear need to conduct hazardous materials characterization and assessment as a prerequisite in material selection and waste stream prioritization for commercial and industrial applications. For this, there has been significant previous work to assess what makes materials hazardous or "less green" through characterization of the many environmental and health hazard traits they present.

However, even with substantial progress on hazardous materials characterization and assessment frameworks, significant problems with current approaches still remain to be resolved. Of these issues, the first is that, internationally, toxicity data and classification systems on hazard traits are not uniform, leading to diverse methodologies to interpret them. Future efforts should move toward a unified toxicity data system (such as GHS), which can ultimately enable standardization of screening-level and impact-pathway assessments of materials because they will be based upon the same toxicological foundation. This will also help reduce the complexity of gathering data from dissimilar toxicity data sources to be used in hazard assessments. A second issue is that material life cycle hazards are usually not reflected in current hazardous material assessment frameworks. Material selection processes that do not consider implications of downstream and upstream chemicals involved in the production and end-of-life management of the material may not meet the objective of true hazard reduction, which is, in reality, a life cycle consideration. The last, perhaps obvious, problem is that in order to motivate the selection of benign materials for use in industry, these assessment tools must be put into use by practitioners. Responsible materials use relies on companies to eventually adopt these methods as a fundamental component in the design process. To this extent, hazard assessment tools must be tailored in a way that best fits the needs of the engineer and designer.

As a final note, benchmarking of material hazard traits that can adversely affect environmental and human health is essentially only one factor to consider (albeit a complex one due to the various hazard endpoints in question) when attempting to achieve the overall objective of sustainable materials use. Environmental trade-off considerations outside of the scope of hazardous materials characterization and assessment such as materials embodied energy use, recyclability, and resource depletion potential should be investigated in conjunction to promote sustainable material choices in global industries. At some point in the future, research effort, specifically in the arena of industrial ecology, needs to assimilate material hazard assessment methods into the larger framework of design for environment strategies. This ideology can ultimately help society move toward a more holistic approach to manage hazardous materials [81].

Bibliography

Primary Literature

1. Wilson MP, Schwarzman MR (2009) Toward a new U.S. chemicals policy: rebuilding the foundation to advance new science, green chemistry, and environmental health. Environ Health Perspect 117:1202–1209
2. Christensen FM, Olsen IS (2004) The potential role of life cycle assessment in regulation of chemicals in the European Union. Int J Life Cycle Assess 9:327–332
3. Holzinger K, Knill C, Sommerer T (2008) Environmental policy convergence: the impact of international harmonization, transnational communication, and regulatory competition. Int Organ 62:553–587
4. Graedel TE, Allenby BR (1996) Design for environment. Prentice Hall, New Jersey
5. Schoenung JM (2008) Lead compounds. In: Shackelford J, Doremus R (eds) Ceramic and glass materials. Springer, USA, pp 151–167
6. Centers for Disease Control and Prevention (2007) Toxicological profile for lead. U.S. Department of Health and Human Services, Atlanta, GA
7. Nigg JT, Knottnerus GM, Martel MM, Nikolas M, Cavanagh K, Karmaus W, Rappley MD (2008) Low blood lead levels associated with clinically diagnosed attention-deficit/hyperactivity disorder and mediated by weak cognitive control. Biol Psychiatry 63:325–331
8. Filippelli GM, Laidlaw MA (2010) The elephant in the playground: confronting lead-contaminated soils as an important source of lead burdens to urban populations. Perspect Biol Med 53:31–45
9. Mielke HW, Reagan PL (1998) Soil is an important pathway of human lead exposure. Environ Health Perspect 106:217–229
10. Annest JL, Pirkle JL, Makuc D, Neese JW, Bayse DD, Kovar MC (1983) Chronological trend in blood lead levels between 1976 and 1980. N Engl J Med 308:1373–1377
11. United Nations (2009) Globally harmonized system of classification and labelling of chemicals (GHS), 3rd edn. United Nations Economic Commission for Europe, Geneva
12. Chiou CT, Freed VH, Schmedding DW, Kohnert RL (1977) Partition coefficient and bioaccumulation of selected organic chemicals. Environ Sci Technol 11:475–478
13. Veith GD, DeFoe DL, Bergstedt BV (1979) Measuring and estimating the bioconcentration factor of chemicals in fish. J Fish Res Board Canada 36:1040–1048
14. OSHA (1996) Hazard communication standard 1910.1200. U.S. Department of Labor, Washington, DC
15. ACGIH (2009) Guide to occupational exposure values. ACGIH, Ohio
16. NIOSH (2010) Carcinogen list. U.S. Department of Health and Human Services, Atlanta, GA
17. German Research Foundation (2010) Senate commission for the investigation of health hazards of chemical compounds in the work area. German Research Foundation, Bonn
18. World Health Organization (2010) Agents classified by the IARC monographs, vols 1–100. International Agency for Research on Cancer, Lyon
19. European Council (1967) Commision directive 67/548/EEC: approximation of laws, regulations and administrative provisions relating to the classification, packaging and labeling of dangerous substances. Brussels, Belgium, European Union
20. NTP (2005) 11th report on carcinogens. U.S. Department of Health and Human Services, Atlanta, GA
21. USEPA (1991) Guidance for water quality-based decisions: the TMDL process, 440/4-91-001. U.S. Environmental Protection Agency, Washington, DC
22. Electronic Code of Federal Regulations (2011) Title 40 protection of environment: part 131.38 establishment of numeric criteria for priority toxic pollutants for the State of California. U.S. Government Printing Office, Washington, DC
23. German Institute for Occupational Safety and Health (2010) Database on hazardous substances. German Institute for Occupational Safety and Health, Sankt Augustin
24. Lodgson MJ, Hagelston K, Mudder TI (1999) The management of cyanide in gold extraction. International Council on Metals and the Environment, Ontario
25. Veeken AH, Rulkens WH (2003) Innovative developments in the selective removal and reuse of heavy metals from wastewaters. Water Sci Technol 47:9–18

26. World Health Organization (1997) IARC monographs on the evaluation of carcinogenic risks to humans: 1,3-butadiene, ethylene oxide and vinyl halides (vinyl fluoride, vinyl chloride and vinyl bromide). 97. International Agency for Research on Cancer, Lyon
27. Allen MR, Braithwaite A, Hills CC (1997) Trace organic compounds in landfill gas at seven U.K. waste disposal sites. Environ Sci Technol 31:1054–1061
28. Rushton L (2003) Health hazards and waste management. Br Med Bull 68:183–197
29. Japanese National Institute of Technology and Evaluation (2010) Information about the status of the implementation of GHS in Japan, chemical management field. Japanese National Institute of Technology and Evaluation, Shibuya-ku, Tokyo
30. Brydson J (2000) Plastics materials, 7th edn. Butterworth-Heinemann, MA
31. Harper CA (2000) Modern plastics handbook. McGraw-Hill, New York
32. Zweifel H, Maier RD, Schiller M (2009) Plastics additive handbook, 6th edn. Hanser Publications, New York
33. Clean Production Action (2010) Plastics Scorecard. Clean Production Action, Seattle, WA
34. Cusack P, Perrett T (2006) The EU RoHS directive and its implications for the plastics industry. Plast Addit Compd 8:46–49
35. Krishnan AV, Stathis P, Permuth SF, Tokes L, Feldman D (1993) Bisphenol-A: an estrogenic substance is released from polycarbonate flasks during autoclaving. Endocrinology 132:2279–2286
36. Yamamoto T, Yasuhara A (1999) Quantities of bisphenol A leached from plastic waste samples. Chemosphere 38: 2569–2576
37. Pearson SD, Trissel LA (1993) Leaching of diethylhexyl phthalate from polyvinyl chloride containers by selected drugs and formulation components. Am J Health Syst Pharm 50:1405–1409
38. Jaakkola JJ, Knight TL (2008) The role of exposure to phthalates from polyvinyl chloride products in the development of asthma and allergies: a systematic review and meta-analysis. Environ Health Perspect 116:845–853
39. Menad N, Björkman B, Allain EG (1998) Combustion of plastics contained in electric and electronic scrap. Resour Conserv Recy 24:65–85
40. USEPA (2011) Dioxins and Furans, Persistent Bioaccumulative and Toxic (PBT) Chemical Program. U.S. Environmental Protection Agency, Washington, DC
41. Goyer RA (1997) Toxic and essential metal interactions. Annu Rev Nutr 17:37–50
42. Barnhart J (1997) Chromium chemistry and implications for environmental fate and toxicity. J Soil Contam 6: 561–568
43. World Health Organization (1997) IARC monographs on the evaluation of carcinogenic risks to humans: chromium, nickel and welding 49. International Agency for Research on Cancer, Lyon, France
44. USEPA (2007) Framework for metals risk assessment, 120/R-07/001. U.S. Environmental Protection Agency, Washington, DC
45. Bridge G (2004) Contested terrain: mining and the environment. Annu Rev Environ Resour 29:205–259
46. Moore JN, Luoma SN (1990) Hazardous wastes from large-scale metal extraction: a case study. Environ Sci Technol 24:1278–1285
47. Kendig M, Jeanjaquet S, Addison R, Waldrop J (2008) Role of hexavalent chromium in the inhibition of corrosion of aluminum alloys. Surf Coat Technol 140:58–66
48. Thomas VM, Graedel TE (2003) Research issues in sustainable consumption: toward an analytical framework for materials and the environment. Environ Sci Technol 37:5383–5388
49. Lim SR, Lam CW, Schoenung JM (2010) Quantity-based and toxicity-based evaluation of the U.S. toxics release inventory. J Hazard Mater 178:49–56
50. Lam CW, Lim SR, Schoenung JM (2011) Environmental and risk screening for prioritizing pollution prevention opportunities in the U.S. printed wiring board manufacturing industry. J Hazard Mater 189:315–322
51. Kurk F, Eagan P (2008) The value of adding design-for-the-environment to pollution prevention assistance options. J Clean Prod 16:722–726
52. Kuczenski B, Geyer R (2010) Chemical alternatives analysis: methods, models and tools. California State Department of Toxic Substances Control, Sacramento, California
53. Saur K (1997) Life cycle impact assessment. Int J Life Cycle Assess 2:66–70
54. Bare JC, Norris GA, Pennington DW, McKone TE (2002) The tool for the reduction and assessment of chemical and other environmental impacts. J Ind Ecol 6:49–78
55. Zhou X, Schoenung J (2009) Combining U.S.-based prioritization tools to improve screening level accountability for environmental impact: the case of the chemical manufacturing industry. J Hazard Mater 172:423–431
56. Lim SR, Kang D, Ogunseitan OA, Schoenung JM (2011) Potential environmental impacts of light-emitting diodes (LEDs): metallic resources, toxicity, and hazardous waste classification. Environ Sci Technol 45:320–327
57. Lim SR, Schoenung JM (2010) Toxicity potentials from waste cellular phones, and a waste management policy integrating consumer, corporate, and government responsibilities. Waste Manag 30:1653–1660
58. Lim SR, Schoenung JM (2010) Human health and ecological toxicity potentials due to heavy metal content in waste electronic devices with flat panel displays. J Hazard Mater 177:251–259
59. Thorneloe SA, Weitz K, Jambeck J (2007) Application of US decision support tools for materials and waste management. Waste Manag 27:1006–1020

60. Hertwich EG, Mateles SF, Pease WS, McKone TE (2001) Human toxicity potentials for life-cycle assessment and toxics release inventory risk screening. Environ Toxicol Chem 20:928–939
61. Lawrence Livermore National Lab (1993) CalTOX, a multimedia total exposure model for hazardous-waste sites. UCRL-CR-111456 (I), Livermore, California
62. Bare JC (2011) TRACI 2.0: the tool for the reduction and assessment of chemical and other environmental impacts 2.0. Clean Technologies and Environmental Policy (in press) doi:10.1007/s10098-010-0338-9
63. Rosenbaum RK, Bachmann TM, Gold LS, Huijbregts MA, Jolliet O, Juraske R, Koehler A, Larsen HF, MacLeod M, Margni M, McKone TE, Payet J, Schuhmacher M, van de Meent D, Hauschild MZ (2008) USEtox – the UNEP-SETAC toxicity model: recommended characterisation factors for human toxicity and freshwater ecotoxicity in life cycle impact assessment. Int J Life Cycle Assess 13:532–546
64. Huijbregts MA, Hauschild MZ, Jolliet O, Margni M, McKone TE, van de Meent D (2010) USEtox^{TM} user manual. USEtox^{TM} team
65. USEPA (2010) Risk-screening Environmental Indicators (RSEI) methodology: RSEI version 2.3.0. Washington, DC
66. Toffel MW, Marshall JD (2004) Improving environmental performance assessment: a comparative analysis of weighting methods used to evaluate chemical release inventories. J Ind Ecol 8:143–172
67. Nissen NF, Griese I, Middendorf A, Müller J, Pötter H, Reichl H (1998) An environmental comparison of packaging and interconnection technologies. Proceedings of the 1998 IEEE international symposium electronics and the environment. Institute of Electrical and Electronics Engineers, New York City, NY
68. Middendorf A, Nissen NF, Griese H, Muller J, Potter H, Reichl H, Stobbe I (2000) EE-toolbox-a modular assessment system for the environmental optimization of electronics. Proceedings of the 2000 IEEE international symposium electronics and the environment. Institute of Electrical and Electronics Engineers, New York City, NY
69. Griese H, Müller J, Reichl H, Zuber KH (2006) Sustainable development of microelectronics industry the role of metals in future interconnection technologies. Sust Metals Manag 19:577–592
70. Yen SB, Chen JL (2009) Calculation of a toxic potential indicator via Chinese-language material safety data sheets. J Ind Ecol 13:455–466
71. Schneider F, Salhofer S, Schiffleitner A, Stachura M (2008) The development of an ecodesign product – the ecomouse case study. Prog Ind Ecol 5:102–123
72. European Council (2001) Commission directive 2001/59/EC annex III: nature of special risks attributed to dangerous substances and preparations. Brussels, Belgium, European Union
73. Fujino M, Suga T, Hamano H (2005) Customization of the toxic potential indicator for Japanese regulation. Fourth international symposium on environmentally conscious design and inverse manufacturing. Tokyo, Japan
74. Swanson MB, Davis GA, Kincaid STW, Bartmess JE, Jones SL, George EL (1997) A screening method for ranking and scoring chemicals by potential human health and environmental impacts. Environ Toxicol Chem 16:372–383
75. Greitens TJ, Day E (2007) An alternative way to evaluate the environmental effects of integrated pest management: pesticide risk indicators. Renew Agr Food Syst 22:213–222
76. Dunn A (2009) A relative risk ranking of selected substances on Canada’s National Pollutant Release Inventory. Hum Ecol Risk Assess 15:579–603
77. Clean Production Action (2009) The Green Screen for Safer Chemicals version 1.0. Clean Production Action, Seattle, WA
78. Betts KS (2007) Formulating green flame retardants. Environ Sci Technol 41:7201–7202
79. USEPA (2002) Design for the environment program fact sheet, 744-F-00-019
80. Maffini MV, Rubin BS, Sonnenschein C, Soto AM (2006) Endocrine disruptors and reproductive health: the case of bisphenol-A. Mol Cell Endocrinol 254–255:179–186
81. Geiser K (2001) Materials matter. MIT Press, Cambridge, MA

Books and Reviews

Andrady AL (2003) Plastics and the environment. Wiley, New Jersey

Bare JC (2010) Life cycle impact assessment research developments and needs. Clean Technol Environ Policy 12: 341–351

Commoner B (1997) The relation between industrial and ecological systems. J Clean Prod 5:125–129

Graedel TE, Allenby BR (2010) Industrial ecology and sustainable engineering. Prentice Hall, New Jersey

Kuczenski B, Geyer R, Boughton B (2011) Tracking toxicants: toward a life cycle aware risk assessment. Environ Sci Technol 45:45–50

Landner L, Reuther R (2004) Metals in society and in the environment. Kluwer, Dordrecht

Lavoie ET, Heine LG, Holder H, Rossi MS, Lee RE, Connor EA, Vrabel MA, DiFiore DM, Davis CL (2010) Chemical alternatives assessment: enabling substitution to safer chemicals. Environ Sci Technol 44:9244–9249

McDonough W, Braungart M (2002) Cradle to cradle: remaking the way we make things. North Point Press, New York

Pittinger CA, Brennan TH, Badger DA, Hakkinen PJ, Fehrenbacher MC (2003) Aligning chemical assessment tools across the hazard-risk continuum. Risk Anal 23:529–535

Ramani K, Ramanujan D, Bernstein WZ, Zhao F, Sutherland J, Handwerker C, Choi JK, Kim H, Thurston D (2010) Integrated sustainable life cycle design: a review. J Mech Des 132:1–15

Shafer DA (2006) Hazardous materials characterization: evaluation methods, procedures, and considerations. Wiley, New Jersey

Hazardous Waste Incineration Ashes and Their Utilization

L. Reijnders
Institute for Biodiversity and Ecosystem Dynamics,
University of Amsterdam, Amsterdam,
The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
Hazard and Ashes
Utilization of Hazardous Waste Incineration Ashes
Future Directions
Bibliography

Glossary

Ash Solid incineration residue.

Hazard Potential to do harm.

MSWI Municipal solid waste incinerator.

Risk Chance that harm occurs.

WTE incinerator Waste-to-energy incinerator, incinerator also producing useful energy (e.g., electricity).

Definition of the Subject

This chapter deals with hazardous solid residues, usually called ashes, from waste incineration. What is to be considered hazardous in this context shows geographical and temporal variability. Currently, hazardous waste incineration ashes are mostly dumped, or disposed of, in landfills or ash lagoons. There is however also substantial, but geographically variable, utilization of such ashes, mostly in construction, including civil engineering (e.g., in roads and embankments) and there have been proposals for wider utilization. Much research has been done on better and wider utilization of hazardous waste incineration ashes, but little thereof has found its way to actual commercial practice. At the end of the chapter, the matter will be raised whether current waste incineration can be considered sustainable and which changes may lead to a more sustainable management of wastes that are currently incinerated.

Introduction

Wastes are incinerated on a large scale, which in turn generates solid wastes that are often called ashes. There is a wide variety of wastes that are incinerated, partly solid (such as municipal and hospital solid wastes), partly sludges (such as wastewater treatment, paper, and refinery sludges), and partly liquids [1–13]. Most of the wastes that are incinerated are not considered hazardous on the basis of current regulations. Quantitatively most important among incinerated wastes are municipal wastes originating from households, street cleaning, and commercial and industrial sources [1]. Such wastes can be burned in an uncontrolled way or in municipal waste incinerators (MSWIs). When the latter also supply useful energy (e.g., electricity), they may be called waste-to-energy (WTE) installations. In the European Union, the USA and Japan, municipal waste incinerators generated about 25 Mt (=10^9 kg)/year of combustion ashes by the year 2000 [2]. In Taiwan, the yearly amount of municipal waste incineration ashes was in 2009 ~1.7 Mt [11], and in China, probably in the order of 2–3 Mt, with fly ashes accounting for 0.5–0.8 Mt [12]. There is also substantial dedicated incineration of nonhazardous industrial solid wastes. Depending on the type of industry, such incineration may lead to ashes with contents of hazardous substances that are different from ashes originating in municipal wastes [10].

Another major type of waste that is not usually categorized as hazardous waste and partly incinerated, is wastewater treatment and sewage sludge [3]. For instance, in Japan, about 90% of wastewater treatment and sewage sludges are incinerated [4], whereas by 2007 in the USA about 22% of these sludges were incinerated and in the European Union about 15%. In the European Union, incineration of wastewater treatment and sewage sludge generated about 2 Mt/year of wastewater treatment and sewage sludge incineration ashes in 2007 [2]. In Japan, the amount of such sludge incineration ash was ~0.5 Mt in 2009 and in Taiwan > 0.5 Mt in 2010 [11, 13].

In waste incinerators, there are varying degrees of co-firing with fossil fuels to increase incineration temperature. When there is substantial co-firing of coal, which is the case in municipal waste incinerators in large parts of China and in some sewage sludge incinerators in Europe [14, 15], the composition and characteristics of incineration ashes will be significantly impacted by coal use.

The incineration of "nonhazardous" municipal wastes, wastewater treatment and sewage, and wastewater treatment sludges may generate ashes, which can be considered hazardous according to current regulations [1–3, 5–7, 13]. What is considered hazardous according to these regulations may vary over time and may differ from country to country [1, 7, 16]. The European Union, for instance, currently considers all types of ashes originating in the incineration of "nonhazardous" municipal wastes as "hazardous," contrasting previous practice in several EU countries, whereas in other countries bottom ash from such incinerators may be considered as "nonhazardous" [7].

There is also incineration of wastes that, on the basis of current regulations, are by themselves considered to be hazardous. There is a wide variety of such wastes, ranging from soils heavily contaminated by organic chemicals, several types of discarded preserved wood, and car shredder residues to hospital and radioactive wastes. There are incinerators specializing in burning hazardous wastes, but there is also co-combustion of hazardous waste in power plants and MSWIs. The ashes originating in the incineration of hazardous wastes by specialized incinerators are usually considered hazardous [17, 18], though there also are exceptions. For instance, incineration ashes originating in bone meal and meat from animals suspected to be infected by transmissible bovine spongiform encephalopathy (BSE) are considered to be nonhazardous by the French government [8]. In section Hazard and Ashes of this chapter, the hazard of ashes is considered more closely.

In this chapter, mainly ashes from incinerators that burn wastes in controlled ways will be discussed. Such incinerators may generate a variety of solid wastes. These include grate ash, grate siftings, bottom ash, heat recovery system ash, fly ash, and air pollution control residue. Table 1 gives the definitions of these incineration wastes to the extent that they originate in waste incinerators using grates, energy recovery, and air pollution control systems.

The collection of the various categories of solid incineration wastes mentioned in Table 1 may be different. However, rather often grate ash, grate siftings, and heat recovery system ash are combined and categorized as bottom ash. Fly ash and air pollution control residue may also be combined and categorized together as "air pollution control residues" or "fly ash," comprising all particulates collected from the flue gas of the incinerator before its emission from a smokestack.

The relative amounts of the different categories of solid incinerator wastes outlined in Table 1 can be different. This is illustrated in Table 2 for municipal waste incinerators in the European Union, Japan, and the USA. Table 2 gives ranges for solid incineration wastes per 1,000 kg of incinerated waste.

Publicly available data about the composition and leachability of waste incineration ashes are patchy. Nevertheless, it can be safely stated that the composition of bottom ash, fly ash, and air pollution control residues, as defined in Table 1, may differ much, dependent on the incinerator, including the choice of combustion technology (e.g., rotary kiln, fluidized bed), actual incineration conditions (such as temperature, throughput, waste heterogeneity, and the addition of coal, natural gas, or mineral oil), and the nature of the wastes that are incinerated [6, 19–23]. Large between-incinerator differences occur in the presence of substances relevant to hazard. As (arsenic) concentrations

Hazardous Waste Incineration Ashes and Their Utilization. Table 1 Definitions of solid incineration wastes using grates, energy recovery, and air pollution control systems [3]

Category of solid incineration waste	Definition
Grate ash	Ash collected from grate
Grate siftings	Material collected from underneath the grate
Heat recovery system ash	Ash collected from boiler, superheater, and economizer
Bottom ash	Combination of the previous categories
Air pollution control residues	Particulate material captured from flue gas downstream of air pollution control reagent injection
Fly ash	Particulate matter captured from flue gas prior to the addition of air pollution control reagents

Hazardous Waste Incineration Ashes and Their Utilization. Table 2 Ranges of solid incineration wastes as defined in Table 1 from municipal waste incinerators in kg/1,000 kg waste [140]

Type of solid incineration waste	Bottom ash	Grate siftings	Heat recovery system ash	Air pollution control residues	Fly ash
Amount of solid incineration waste in kg/ 1,000 kg waste	100–250	<1–3	<2–5	10–60	7–25

in bottom ash from municipal waste incinerators may, for instance, vary by over a factor 1,000 and Hg, Pb, Ni, Mo, Cr, and Cd concentrations by over a factor 100 [19]. Differences in chloride and lead concentrations in air pollution control residues, as defined in Table 1, exceed a factor of 10 [23].

Differently from, e.g., a plant manufacturing yogurt, the composition of a specific incineration waste (e.g., bottom ash) from a waste incinerator may vary substantially from day to day. Also, after their generation, waste incineration ashes can undergo major changes due to "aging," which in turn may have a major impact on leachability [2, 5, 24, 25].

Furthermore, there are differences between bottom ash, fly ash, and air pollution control residues originating from the same waste and the same incinerator. Bottom ash is usually enriched in inorganic compounds, which are relatively nonvolatile on incineration, and fly ash and air pollution control residues tend to be enriched in volatile inorganic compounds. It has also been found that bottom ash may be more enriched in adsorbable organohalogens [20] and polycyclic hydrocarbons [18, 26] than fly ash.

In this chapter, the emphasis will be on the utilization of ashes from municipal waste and wastewater treatment and sewage sludge incinerators. Because this encyclopedia deals with sustainability science and technology, in section "Future Directions" the question will be raised whether current waste incineration practices and handling, including utilization of ashes, can be considered sustainable. This raises the contentious matter of intergenerational equity of waste handling [27].

Hazard and Ashes

As pointed out in the introduction, there can be differences between countries as to the categorization of ashes as hazardous. Hazardous may in this context refer to humans and/or organisms in the environment.

When ashes are released to air, as may occur, for instance, in the working environment, and at the end of smoke stacks, the main hazard to humans is linked to fly ash, which contains relatively small particles. The hazard of fly ash particles is related to their size, their surface area, structure, charge, agglomeration, and the presence of specific chemicals [28–30]. Exposure to fly ash adds to the background exposure to small particles in air. In urbanized settings, such background exposure is often associated with substantial risk of cardiovascular and pulmonary disease and exposure to fly ash may increase such risk [28, 29]. Also worker exposure to MSWI fly ash has been associated with increased DNA strand breakage, which may be linked with increased cancer risk [31].

Apart from the emission of ash particles to air, there is substantial agreement that the hazard of ashes is also dependent on the leaching of specific substances.

There have been proposals to use assays measuring the toxicity of leachates to specific organisms ("biotoxicity tests"), which in principle evaluate the complete leachate [32–36]. Such assays may employ strains of bacteria such as *Escherichia coli* [32], *Vibrio fischeri* [15, 33, 34], *Bacillus subtilis* [5], the microcrustacean *Daphnia magna* [15, 35, 36], the alga *Pseudokirchneriella subcapitata* [35, 36], human cell lines [35] and the duckweed *Lemna minor* [36].

A major problem is that there is no proven link between actual risk to humans and/or organisms in the human environment and the outcomes of the proposed biotoxicity tests. This is a common problem with risk assessment of "polluted" materials [37]. Another suggested approach uses an aquatic microcosm, employing a variety of organisms [36, 38]. Though this is more sophisticated than the biotoxicity tests referred to above, the correspondence with effects "in the field," which primarily regards soils, is as yet unknown.

In practice, the hazard of incineration ashes is often regulated on the basis of leaching tests. Usually, a limited number of substances are considered by regulators in this context. These rather often belong to the following categories of substances: relatively toxic elements such as heavy metals and arsenic and the relatively toxic polycyclic and halogenated organic compounds.

A problem is that the substances that are actually leached from ashes may show substantial variations in concentration dependent on conditions such as weathering of ashes, the presence of organic substances such as humic acids and of specific microorganisms, and on factors such as pH, and aerobic or anaerobic ("redox") conditions [6, 39–41].

Standardized leaching tests in the laboratory which underpin estimates of, and regulatory decisions as to, hazard posed by ashes, tend to reflect a limited subset of the conditions as they occur "in the field." This may lead to large differences between leaching as suggested by laboratory tests and actual leaching, which in turn may lead to estimates of hazard and risk that are at variance with actual hazard and risk [2, 40, 42–46].

Whether leached substances should be a matter of concern is dependent on exposure to leached substances and background exposure to the substances involved. For instance, there is evidence that present exposure to metals such as Cd and Pb and to halogenated compounds, such as chlorinated biphenyls, -dioxins, and -benzofurans, and polycyclic hydrocarbons is a health risk to the general human population in industrialized countries [37] and this is a good reason to even consider ashes which leach such substances in very small amounts as hazardous.

Utilization of Hazardous Waste Incineration Ashes

Treatment of Ashes Before Utilization

There has been substantial research into treatment of ashes before utilization. These treatments are aimed at eliminating small metal particles, lowering the amount of hazardous and/or mobile substances and at improving ash performance during utilization in specific applications [2, 42, 47].

A process that has been developed to remove small metal particles (Ag, Cu, Pb, Sn, and Zn), is based on eddy current and magnetic separation [43]. As the particles collected by this process are high-value metals, this treatment may be profitable, depending on actual metal prices.

Other treatments have been developed to improve ash-related product performance during utilization in, or as, building and geotechnical materials. Such treatments include those mentioned in Box 1 [2, 42, 47]:

Many treatments aimed at lowering the amount of hazardous and/or mobile substances have been proposed [2, 42, 47–54]. These can be categorized as indicated in Box 2.

There are also treatments intended to be used prior to utilization, which aim at immobilization of hazardous substances. These include the treatments in Box 3 [2, 10, 11, 17, 24, 43, 49, 54, 55]:

Treatments aiming at immobilization, or lowering the amount, of hazardous substances, may not only precede utilization, but may also serve reduced leaching from landfilled hazardous waste incineration ashes.

There are substantial differences in costs between the pretreatments summarized in Boxes 2 and 3. It would seem that aqueous mobilization techniques tend to be less expensive than thermal mobilization and immobilization technologies [45].

Commercial use of most of the treatments prior to utilization outlined in this section is limited or absent [2, 10, 11, 43, 49]. To the extent that there is treatment, the three treatments in Box 4 are practiced on a large commercial scale. These treatments may be combined [e.g., 11].

Still, much of the utilization of hazardous waste incineration ashes is not preceded by the treatments summarized in Box 4.

Application in or as Building and Geotechnical Materials

To the extent that there is utilization of hazardous waste incineration ashes, most ashes are used in the construction sector, including civil engineering, thus in, or as, building and geotechnical materials [2, 7, 56]. Most of the ashes which are actually applied seem to be bottom ashes, as defined in Table 1, but there is also some application of fly ashes (as defined in Table 1), for instance in asphalt mixtures [2]. Bottom ashes are used in the production of bricks, cement, and

Box 1. Treatments to Improve Ash-Related Product Performance During Utilization in Building and Geotechnical Materials

- Aqueous washing to remove soluble salts
- Oxidation of Al to reduce the generation of H_2 at high pH values
- Thermal treatment to reduce the carbon content of ashes

Box 2. Treatments Aimed at Lowering the Amount of Hazardous and/or Mobile Substances

- Treatments aiming at the destruction of organic (carbonaceous) substances present in ashes. These mainly include thermal, hydrothermal, and thermocatalytic treatment methods. Supercritical fluid treatment to destruct organic substances has also been investigated.
- Treatments aimed at the mobilization of hazardous and/or mobile substances. These include:
  - Aqueous washing to remove easily soluble inorganic compounds such as chlorides and sulfates and fine particulate matter, which is usually characterized by relatively high levels of hazardous substances
  - Supercritical fluid treatment to mobilize heavy metals
  - Electrodialysis for the mobilization of heavy metals, such as Cu, Cr, Zn, and Cd
  - Acid leaching for the mobilization of metals
  - Use of chelating substances for the mobilization of heavy metals
  - Biological mobilization using microorganisms such as *Thiobacillus thiooxidans, Rhodococcus,* and *Aspergillus niger* for the removal of metals
  - Pyrometallurgical or thermochemical mobilization of metals in the presence of excess Cl
  - Use of organic extractants to remove heavy metals

Box 3. Treatments Aiming at Immobilization of Hazardous Substances

- Thermal treatments such as sintering, (co)melting, and calcination.
- Natural weathering including natural carbonation. This results in decrease of pH, mildly reducing conditions, dissolution and precipitation reactions, hydrolysis/hydration, and mineral changes, which reduce leaching of a number of elements (though leaching of some elements (e.g., Cr, Sb) may increase) (see also "Life cycle release of hazardous substances").
- Enhanced or accelerated carbonation.
- Stabilization with phosphate, which may increase the sequestration of metals such as Zn, C, Pb, and Cd.
- Hydrothermal treatment in the presence of additives such as alkaline substances or $FeSO_4$.
- Treatment with supercritical fluids.
- Mechanochemical treatments, such as wet ball milling in the presence of CaO.

Box 4. Treatments of Hazardous Waste Incineration Ash Prior to Potential Utilization, Practiced on a Large Commercial Scale

- Natural weathering.
- Aqueous washing to remove highly soluble and mobile substances and fine particulate fractions that tend to contain relatively high amounts of hazardous substances from ashes [2].
- Melting for immobilization [10, 11, 49]. By the end of 2004, 140 plants melting waste incineration ashes were operational in Japan [49].

cement-derived products such as concrete, in sublayers of roads, cycling tracks, and parking spaces, in covers of landfills, and in embankments and noise insulation works [2, 5]. By the year 2000, in Germany, about 50% of bottom ash from municipal waste incineration was used in this way, in the Netherlands 60% and in Denmark 72% [5]. In China, substantial use of waste incineration ashes in road construction and brick making has been reported [53]. In the USA and

Canada, little actual utilization of hazardous waste incineration ashes seems to occur [5]. The same seems to hold for several other countries [57–60]. There are no good quantitative data regarding the application of hazardous waste incineration ashes per category of building and geotechnical materials, but it would seem that most of the ashes are applied in civil engineering: as or in geotechnical materials.

Utilization of waste incineration ashes can have both positive and negative environmental effects [37]. Utilization of waste incineration ashes in buildings and civil engineering may substitute for virgin resources, which is commonly considered a beneficial environmental effect [61]. On the other hand, hazardous substances may be released. These may be associated with risk and also with detrimental effects on natural resources such as groundwater [61–65]. For instance, in a leaching study of asphalt pavement amended with "waste-to-energy" ash, it was found that aluminum concentrations in leachates were consistently above water quality standards applicable in Florida [64]. In a life cycle assessment study comparing use of "upgraded" bottom ash in Swedish road construction with landfilling of bottom ash, the difference in environmental impact was found to be marginal [61].

It may be noted too, that some types of treatment of incineration ashes before application, are associated with large inputs of energy and/or materials. Such treatment may be a significant determinant of environmental impact [66]. For instance, utilization options involving thermal treatment may be very energy intensive, which may represent a major environmental burden [66].

Life Cycle Release of Hazardous Substances The release of hazardous substances during the life cycle of building and geotechnical materials including hazardous waste incineration ashes has been studied to a limited extent. Firstly, the release of hazardous substances may originate in the production of building materials and the realization of the building or engineering work. There may be leaching from the intact product or civil engineering work, and following wear and tear during the use stage of the life cycle. The latter is for instance prominent in the case of asphalt concrete aggregates applied in areas with rather severe winters such as Scandinavia, Russia, parts of Japan, and the USA, where antiskid materials are applied during the winter. This gives rise to large releases of fine particulate road dust (from which substances may leach) in spring and winter [67]. Leaching may also be impacted by handling after the use stage (e.g., during demolition or reuse) [2].

Knowledge about the release of hazardous substances originating in incineration ashes during the life cycle of buildings and civil engineering work is patchy and mostly concerns ashes that have not been subject to pretreatment [2, 3, 42, 51, 52, 63, 68–71].

Most work on the release of hazardous substances published in peer-reviewed journals has been done on the utilization of ashes in the production of cement and cement clinker. In view of the very high temperature and relatively long residence time involved in clinker production, the input of hazardous waste incineration ashes in cement clinker production can be instrumental in high-efficiency break down of toxic organics [72]. Adequate preventive measures should be taken against *de novo* formation of toxic compounds, such as low scrubber inlet temperatures to prevent the formation of chlorinated dioxins [73]. In actual practice the *de novo* formation of toxics may be significant however. For instance, in Taiwan, it has been estimated that on including hazardous residues in cement production, the emission of chlorinated dioxins may increase from 0.29 ng I-TEQ (international toxic equivalents)/kg clinker produced to 24.34 ng I-TEQ/kg [73]. In the case that ashes are included in cement clinker production, there tend to be significant stack emissions of heavy metals from the cement factory [64]. The incorporation of waste incineration ashes in cement clinker and cement may lead to substantial leaching of hazardous compounds during the life cycle of cement-derived products, including the recycling stage [70]. Hazardous elements present in cement or concrete will not pollute when they remain immobile. Two factors contribute to the immobilization of hazardous elements: high pH and inaccessibility to the solvent. In practice, these protective factors do not allow for complete immobilization of all hazardous elements during the life cycle of building and geotechnical cement-derived products including hazardous waste incineration ashes. Oxyanions such as CrO_4^{2-}, AsO_4^{3-}, MoO_4^{2-}, SeO_4^{2-}, and VO_4^{3-} are mobile at high pH [74]. Actual leaching of these anions from mortars shows high variability between producers. In a study of 56 cement mortars of worldwide origin, van der Sloot et al. [75] found, for instance, that the leaching of vanadium varied by about two orders of magnitude. The mobility of oxyanions from mortars may have health implications for workers. It has been found that elevated concentrations CrO_4^{2-} in cement that is handled by workers may lead to allergic skin complaints [76]. This has led to legislation, first in the Scandinavian countries [77] and more recently to a Directive of the European Union (EU), which limits the presence of soluble CrO_4^{2-} in cement [78]. Also, CrO_4^{2-} is carcinogenic to humans after inhalation [79]. This is also relevant to the health of workers handling cement [78]. Furthermore leaching of CrO_4^{2-} from cement products may exceed water quality standards [e.g., 65]. One strategy to meet the limitation to soluble Cr^{6+} under EU legislation is to add ferrous sulfate, to reduce Cr^{6+} to Cr^{3+} [76]. This strategy has limited applicability, so there is substantial interest in restricting the input of materials with relatively large concentrations of soluble Cr^{6+} into cement kilns [76]. Metal chlorides may also be incompletely immobilized in cement and cement-derived products [2].

Once incorporated into concrete or other cement-derived products, the long-term fate of hazardous elements present in cement is a matter to consider. Firstly, there is the matter of mobilization from the product during use. Guo [80] has pointed out that use of cement mortar lined pipes for the transport of drinking water may necessitate limits on the presence of chromium and lead in inputs, in view of leaching from the pipes of $CrO4^{2-}$ and Pb. Metal chlorides may be present in high concentration in fly ash from waste incinerators, when there is a high input of materials which generate HCl gas when incinerated, such as polyvinylchloride (PVC) [81]. When clinker is made from such municipal waste incineration ash, substantial leaching of metal chlorides from cement-derived products may occur [82, 83]. The fate of hazardous elements present in cement-derived products during use is dependent on the processes that affect such products and environmental factors. These may include wear and tear, cracking, weathering including carbonation (which lowers pH), leaching, exposure to humic acids (that may increase metal solubility), and microbial activity. Wear and tear, cracking, weathering, and leaching may increase the interaction of solvent with hazardous elements. It has furthermore been shown in laboratory experiments that leaching of elements such as As, Cr, Ni, Sb, and Zn from cement may in the long run increase because of carbonation [83–87]. Microbial activity, according to laboratory experiments, may lead to the mobilization from cement-derived products of Cr [88] and probably also of Pb, Cu, and Zn [89, 90].

It is hard to predict the extent to which processes such as weathering, carbonation, leaching, exposure to humic acids or microorganisms, and wear and tear will lead to actual mobilization of hazardous elements from cement-derived products. As pointed out before, the standard leaching tests as they have been developed for laboratory use do not take account of microbial activity nor of the potential effect of humic acids which may increase mobility. They are also inadequate in representing weathering [2, 40, 42–46, 91–94]. This may lead to predictions of leaching on the basis of tests that are much at variance with actual leaching [40, 42–46, 92, 94]. However, all in all, it may well be that in the long run substantial amounts of hazardous elements will leach from waste-based cement-derived products in actual use.

When concrete structures are demolished at the end of their service life, leaching may become more of a problem. Van der Sloot [95] has shown that, whereas the leachability of Cd, Cr, Mo, and V is low in freshly cured concrete, leaching from construction debris used as an aggregate, e.g., in road stabilization bases, is much increased and may well exceed the statutory leaching limits as they are used in the Netherlands. Serclérat et al. [96] have shown that after incorporation in Portland cement clinker, much CrO_4^{2-} is retained in intact mortar bars leached with demineralized water, but that practically all CrO_4^{2-} is extracted from crushed material, which has a much increased surface area. This implies that without additional containment, much of the CrO_4^{2-} that was meant to be permanently immobilized in concrete may be released into the environment during the life cycle of cement products. A similar fate may apply to other mobile oxyanions of hazardous elements and to metal chlorides.

In view of substantial leaching of hazardous substances from utilized ashes, there would seem to be a case for pretreatments of incineration ash that would lead to the removal of hazardous compounds

[70]. These pretreatments were outlined in Boxes 1 and 2 of section "Treatment of Ashes Before Utilization". Also there would seem to be a case to explore the options for source reduction and separation that would lead to a reduction of leachable hazardous substances. This applies both to hazardous elements and organics, dependent on the processing of ashes. As to the latter, Hunsinger et al. [97–99] have proposed real-time automatic control of primary air supply, grate kinematics and waste feeding rate, and SO_2 recycling to improve burnout, which would lower the concentration of polycyclic aromatics, and to reduce chlorinated dioxin and benzofuran levels in MSWI ashes by about an order of magnitude. The reduction of polychlorobiphenyl levels in MSWI ashes by SO_2 recycling may be more limited than the reduction of chlorinated dioxin levels [100]. As to hazardous elements, for instance, reduced inputs of Pb, As, and Mo in municipal waste incinerators may reduce the potentially leachable amounts of these elements in waste incineration ashes [101, 102]. Similarly, reduced discharges of heavy metals into the sewer system may diminish the presence thereof in ashes originating in the incineration of wastewater treatment and sewage sludges [103]. Lowering the input of heavy metals may in turn, *ceteris paribus,* also lead to lower concentrations of hazardous halogenated compounds in fly ashes [104].

Mobilization of Useful Substances

Waste incineration ashes may contain substantial amounts of useful substances. In this respect attention has been drawn to the presence of substances derived from geochemically scarce natural resources that are formed in slow geological processes. Such substances include phosphate and metals, except Fe and Al. A technology based on eddy current and magnetic separation to collect small particles with geochemically scarce metals has already been referred to in section "Treatment of Ashes Before Utilization". Phosphate has also drawn attention. Sewage and wastewater treatment sludge incineration ashes in European countries have been found to contain 9–21% of P, specified as P_2O_5 [105, 106]. Several methods have been proposed for the extraction of phosphate from such ashes, including extraction with acid [107–109]. Thermochemical treatment leading to the removal of heavy metals and increased biological availability of phosphate present in ashes has also been proposed [105, 106]. The latter ties in with thermochemical or pyrometallurgical treatments in the presence of added chlorine for the winning of heavy metals [50, 105, 106]. This was one of the options mentioned in Box 2 of section "Treatment of Ashes before Utilization", which dealt with the pretreatment of ashes before utilization.

Box 5. Options for the Extraction of Metals from Waste Incineration Ashes

- Supercritical fluid extraction
- Electrodialysis
- Acid leaching
- Use of chelating substances
- Biological mobilization using microorganisms such as *Thiobacillus thiooxidans*, *Rhodococcus,* and *Aspergillus niger*
- Use of organic extractants

Other options to extract metals from waste incineration ashes are in Box 5 [2, 42, 47–51, 110].

There is as yet, no significant commercial exploitation of the technologies for mobilization of phosphate and the extraction of metals by the processes mentioned in Box 5 or pyrometallurgical treatments. However, Germany and Sweden have announced that they will set targets for the recycling of phosphates that currently end up in sewage and wastewater treatment sludges [111]. When recycling targets are implemented, phosphate extraction from ashes will have to compete with phosphate extraction from the original sludges.

Application in Agriculture

There have been proposals for the direct application of wastewater treatment and sewage sludge incineration ashes in agriculture, mainly in view of the presence of phosphate [111, 112]. However, the reason that such sludges are incinerated is usually that they are considered unfit for application in agriculture in view of the presence of substances that can accumulate in soils such as heavy metals [108, 110, 111, 113]. The latter also end up to a large extent in ashes [110, 111, 113]. Moreover, the biological availability of phosphate in sludge incineration ashes tends to be relatively poor, if compared with the original sludges [106, 107, 109]. No substantial utilization of wastewater treatment and sewage sludge incineration ashes seems to occur as yet.

Also there have been proposals for the application of MSWI ashes in agriculture as soil amendment [112]. This may lead to elevated concentrations of heavy metals such as Pb and Cd in crops cultivated on the amended soil [114–116]. So far, there seems to be no substantial utilization of municipal waste incineration ashes generated in a controlled way, for this purpose.

In urban agriculture in Nigeria, town refuse ash, generated by uncontrolled incineration, is used, which aims at reducing soil acidity and providing for a range of micronutrients [117]. Samples of such ash have been studied to a limited extent. It has been found that their composition is very variable, with some samples being strongly contaminated with heavy metals [118].

Applications in Wastewater Treatment

There have been proposals to apply waste incineration ashes, and products derived from those ashes in wastewater treatment. These include:

- Application in sludge conditioning
- Use as package material for anaerobic filters
- Use of incineration ash and incineration ash-derived materials as adsorbent for removal of substances from wastewater

Sludge Conditioning Wastewater treatment sludges should be dewatered before further processing. Such dewatering may be difficult due to the presence of oil. To facilitate dewatering oily sludge, conditioners can be added. It has been found that fly ash from municipal waste incinerators (at an optimum dosage of 3%) can facilitate dewatering by filtration [119]. In doing so, such ashes may add heavy metals to the discharge of the wastewater treatment plant [119]. So far, there seems to be no substantial utilization of municipal waste incineration fly ash for this purpose.

Material for Filters Yue et al. and Han et al. [120, 121] have proposed the use of ceramic particles made from incinerated sludge in filters used for wastewater treatment. Data about the fate of hazardous components of the sludge-derived ceramic particles or the emission of hazardous substances via the filters are absent from their publications [120, 121].

Use as Adsorbent for Removal of Substances from Wastewater Pan et al. [122] and Bouzid et al. [123] have proposed the use of sewage sludge ash to absorb copper from wastewater. Pan et al. [122] have claimed an efficiency of $>98\%$. Hydrothermally treated paper sludge ash [124] and zeolites produced from paper sludge ash [125] or municipal incinerator fly ash [126], untreated bottom ash from MSWIs [127], and medical waste incinerator fly ash, hydrothermally treated with addition of sodium carbonate [128], have also been proposed as adsorbents for (waste) water treatment. Information about adsorption kinetics and mass transfer in these studies [122–128] is very limited. The papers advocating the use of ashes, and ash-derived materials, as adsorbents [122–128] furthermore do not give data about the net effect of such use on the emission of hazardous substances from wastewater treatment plants.

One would expect that the removal of substances from wastewater by ashes would significantly depend on the presence of residual C and would be variable as the composition of such ashes is also variable [129]. Also one would expect that ash-derived adsorbents would be less effective on a kg basis than high-value absorbents currently available [129] and, in line with section "Introduction", that ash-derived adsorbents would be more variable in performance than current high-value adsorbents. It should furthermore be noted in this context that the utility as adsorbents of sewage sludge incineration ash derivatives is less than that for some other sewage-sludge-derived adsorbents, such as sewage sludges treated with alkali metal hydroxide reagents [130]. In practice, there would seem to be, as yet, no substantial applications of incineration ashes in wastewater treatment.

Other Uses as Adsorbent

Uses of hazardous waste incineration ashes in dry processes have also been suggested. MSWI bottom ash has been proposed as a sorbent for the removal of CO_2 from air [131] and of H_2S and CO_2 from biogas produced from agricultural residues and crops [132]. The utilization of hazardous waste incineration ash, possibly combined with coke, has also been suggested for the removal of H_2S and mercaptanes from landfill gas [133]. The use of MSWI ashes for the removal of other volatile organics, HCl, NO_x, and $HgCl_2$ in dry processes might also be considered [133–135]. The studies suggesting application of ashes as adsorbent in dry processes have not looked at the fate of hazardous substances beyond the stage of utilization. There seems to be as yet no substantial commercialization of this type of applications of hazardous waste incineration ashes, a phenomenon that may be linked to better removal efficiencies and a more constant performance achieved by current commercial removal technologies.

Future Directions

In the previous sections, the focus has been on the utilization of hazardous waste incineration ashes. From a sustainability perspective, one may wonder, however, whether incineration, from which the ashes discussed here are derived, is actually always a good thing. Here, sustainability is used in line with its original definition, which means that the practice can be continued indefinitely while keeping natural resources in a steady state, without generating pollution that negatively affects future generations and while keeping ecosystem services at their current level [27, 136]. How does current incineration perform in this respect, if compared with its alternatives?

It would seem that only in a very limited number of cases waste incineration is associated with the best environmental performance. Examples thereof probably include the incineration of infectious wastes and of a variety of hazardous organic chemical wastes, such as some organic solvent wastes [137].

However, for many wastes other options than incineration would seem to be preferable. Municipal waste incineration is currently largely fuelled by synthetic polymers, such as plastics. But for these polymers reuse of product or material is often environmentally preferable [138]. State-of-the-art burning of garden and food wastes for electricity production apparently does not displace fossil fuel use in state-of-the-art power plants [139], which is in favor of conversion methods that allow for the recycling of nutrients and carbon compounds to agricultural land and gardens to keep levels of soil carbon compounds and nutrients in a steady state [136].

Source reduction and separation of hazardous substances combined with agricultural application of source-separated nutrients [113] may well be environmentally preferable to current sewerage and wastewater treatment practices, including sludge incineration.

The choice in favor of incineration is often defended on the basis of cost. But, in current calculations of such costs, many costs relevant to sustainability are largely externalized. Such currently externalized costs include replacement costs for depleted natural resources, damage due to pollution associated with incineration and reduction of ecosystem services, assuming a zero discount, as is proper in view of sustainability [136]. Inclusion of such costs may well lead to very different conclusions about the financial benefits of incineration.

Switching to other options of generating and handling what is currently considered waste will have large implications for the "reverse" logistics and handling of what is disposed of. Such a change should not come as a surprise. There is a very large asymmetry between the complicated and sophisticated system of production and handling products and the relatively uncomplicated and often unsophisticated way in which wastes are currently treated.

Bibliography

1. Johannessen KM (1996) The regulation of municipal waste incineration ash: a legal review and update. J Hazard Mater 47:383–393
2. Reijnders L (2005) Disposal, uses and treatments of combustion ashes: a review. Resour Conserv Recycl 43:313–336
3. Cyr M, Coutand M, Clastres PJ (2007) Technological and environmental behavior of sewage sludge ash (SSA) in cement-based materials. Cem Concr Res 37:1278–1289
4. Murakami T, Suzuki Y, Nagasawa H, Yamamoto T, Koseki T, Hirose H, Okamoto S (2009) Combustion characteristics of sewage sludge in an incineration plant to energy recovery. Fuel Process Technol 90:778–783
5. Sawell SE, Chandler AJ, Eighmy TT, Hartlen J, Hjelmar O, Kosson D, van der Sloot HA, Vehlow J (1995) An international perspective on the characterization and management of residues from MSW incinerators. Biomass Bioenergy 9:377–386
6. Römbke J, Moser T, Moser H (2009) Ecotoxicological characterization of 12 incineration ashes using 6 laboratory tests. Waste Manage 29:2475–2482
7. Gidarakos E, Petrantonaki M, Anastasadou K, Schramm K (2009) Characterization and hazard evaluation of bottom ash produced from incinerated hospital waste. J Hazard Mater 172:935–942

8. Coutand M, Cyr M, Deydier E, GR Clastres (2008) Characteristics of industrial and laboratory meat and bone meal ashes and their potential applications. J Hazard Mater 150: 522–532
9. Karamalidis AK, Voudrias EA (2009) Leaching and immobilization behavior of Zn and Cr from cement-based stabilization/solidification of ash produced from incineration of refinery oily sludge. Environ Eng Sci 26:81–96
10. Sakanakura H (2005) Diffusion test of 20 kinds of waste molten slags and competitive materials. J Mater Cycles Waste Manage 7:71–77
11. Chiang K, Hu Y (2010) Water washing effects on metals emission reduction during municipal solid waste incinerator (MSWI) fly ash melting process. Waste Manage. doi:10.1016/j.wasman.2009.12.009
12. Tian S, Yu M, Wang W, Wang Q, Wu Z (2009) Investigating the speciation of copper in secondary fly ash by x-ray absorption spectroscopy. Environ Sci Technol 43:9084–9088
13. Donatello S, Tyrer M, Cheeseman CR (2010) EU landfill waste acceptance criteria and EU hazardous waste directive compliance testing of incinerated sewage sludge ash. Waste Manage 20:63–71
14. Nie Y (2008) Development and prospects of municipal solid waste (MSW) incineration in China. Front Environ Sci Eng Chin 2:1–7
15. Barbosa R, Lapa N, Boavida D, Lopes H, Gulyurtlu I, Mendes B (2009) Co-combustion of coal and sewage sludge: chemical and ecotoxicological properties of ashes. J Hazard Mater 170:902–919
16. Wiles CC (1996) Municipal solid waste combustion ash: state-of-the-knowledge. J Hazard Mater 47:325–346
17. Reich J (2003) Slag from hazardous waste incineration: reduction of heavy metal leaching. Waste Manage Res 21:110–118
18. Zhao I, Zhang F, Chen M, Liu Z, Wu DBJ (2010) Typical pollutants in bottom ashes from a medical waste incinerator. J Hazard Mater 173:181–185
19. Chandler AJ, Eighmy TT, Hartlen J, Hjelmar O, Kosson DS, Sawell SE, van der Sloot HA, Vehlow J (1997) Municipal solid waste incinerator residues. Elsevier, Amsterdam
20. Durnusoglu E, Bakoglu M, Karademir A, Kirli L (2009) Adsorbable organic halogens (AOX) in solid residues from hazardous and clinical waste incineration. J Environ Sci Health A 41:1699–1714
21. Neuer-Etscheidt K, Nordsieck HO, Liu Y, Kettrup A, Zimmermann R (2006) PCDD/F and other micropollutants in MSWI crude gas and ashes during plant start-up and shut down processes. Environ Sci Technol 40:342–349
22. Rubli S, Medilanski E, Belevi H (2000) Characterization of total organic carbon in solid residues provides insight into sludge incineration processes. Environ Sci Technol 34:1772–1777
23. Quina MJ, Bordado JC, Quinta-Ferreira RM (2008) Treatment and use of air pollution control residues from MSW incineration: an overview. Waste Manage 28:2097–2121
24. Freyssinet P, Piantone P, Azaroual M, Itard Y, Clozel-Lecloup B, Guyonnet D, Baubron JC (2002) Chemical changes and leachate mass balance of municipal solid waste bottom ash submitted to weathering. Waste Manage 22:159–172
25. Dugenies S, Combrisson J, Casablanca H, Grenier-Loustalot MF (1999) Municipal solid waste incineration bottom ash: characterization and kinetic studies of organic matter. Environ Sci Technol 33:1110–1115
26. Liu Y, Li Y, Li X, Jiang Y (2008) Leaching behavior of heavy metals and PAHs from MSWI bottom ash in a long-term static immersing experiment. Waste Manage 28:1126–1136
27. Okrent D (1999) On intergenerational equity and its clash with intragenerational equity and on the need for policies to guide regulation of disposal of wastes and other activities posing very long-term risks. Risk Anal 19:877–901
28. Dellinger B, D'Alessio AD, D'Anna AD, Ciajolo A, Gullett B, Henry H, Keener M, Lighty J, Lomicki S, Lucas D, Oberdörster G, Pitea D, Suk W, Sarofim A, Smith KR, Stoeger T, Tolbert P, Wyzga R, Zimmermann R (2008) Combustion byproducts and their health effects. Environ Eng Sci 25:1107–1114
29. Shih S, Wang Y, Chang J, Jang J, Kuo F, Wang L, Chang-Chien G (2006) Comparisons of levels of polychlorinated dibenzo-p-dioxins/benzofurans in the surrounding environment and workplace of two municipal solid waste incinerators. J Hazard Mater B 137:1817–1830
30. Jiang J, Oberdörster G, Biswas P (2009) Characterization of size, surface charge, and agglomeration state on nanoparticle dispersions for toxicological studies. J Nanopart Res 11:77–89
31. Chen H, Chen I, Chia T (2010) Occupational exposure and DNA strand breakage of workers in bottom ash recovery and fly ash treatment plants. J Hazard Mater 174:23–27
32. Lin K, Chen B (2006) Understanding biotoxicity for reusability of municipal solid waste incinerator (MSWI) ash. J Hazard Mater A 138:9–15
33. Chou J, Wey M, Liang H, Chang S (2009) Biotoxicity evaluation of fly ash and bottom ash from different municipal solid waste incinerators. J Hazard Mater 168:197–202
34. Takeuchi M, Kawahata H, Gupta LP, Itouga M, Sakakibara H, Ohta H, Komai T, Ono Y (2009) Chemistry of fly ash and cyclone ash leachate from waste materials and effects of ash leachates on bacterial growth, nitrogen–transformation activity, and metal accumulation. J Hazard Mater 165:967–973
35. Shoji R, Nakayama H, Sakai Y, Mohri S, Yamada M (2008) Evaluation of the ecotoxicity of solid wastes using rapid leaching test and bioassays. J Environ Sci Health A 43:1048–1053
36. Triffault-Bouchet G, Clement B, Blake G (2005) Ecotoxicological assessment of pollutant flux released from bottom ash reused in road construction. Aquat Ecosyst Health Manage 8:405–414
37. Reijnders L (2009) Are soil pollution risks established by governments the same as actual risks? Appl Environ Soil Sci ID 237038:1–7
38. Clement B, Triffault-Bouchet G, Lottmann A, Carbonel J (2005) Are percolates released from solid waste incineration bottom ashes safe for lentic ecosystems? A laboratory ecotoxicological approach based on 100 litre indoor microcosms. Aquat Ecosyst Health Manage 8:427–439

39. Michalzik B, Ilgen G, Hertel F, Hantsch S, Bilitewski B (2007) Emissions of organo-metal compounds via the leachate and gas pathway from two differently pre-treated municipal waste materials – a landfill study. Waste Manage 27:497–509
40. Sabbas T, Poletti A, Pomi R, Astrup T, Hjelmar O, Mostbauer P, Cappai G, Magel G, Salhofer S, Speiser C, Heuss-Ambicher S, Klein R, Lechner P (2003) Management of municipal solid waste incineration residues. Waste Manage 23:61–88
41. Kim Y, Osako M (2004) Effect of adsorption capacity of dissolved humic matter on leachability of dioxins from raw and treated fly ashes of municipal solid waste incinerators. Arch Environ Contam Toxicol 46:8–16
42. Todorovic J, Ecke H (2006) Treatment of MSWI residues for utilization as secondary construction minerals: a review of methods. Miner Energy 20(4–5):45–59
43. Lidelöw S, Lagerkvist A (2007) Evaluation of leachate emissions from crushed rock and municipal solid waste incineration bottom ash used in road construction. Waste Manage 27:1356–1365
44. Apul D, Gardner K, Eighmy T, Linder E, Frizzell T, Roberson R (2005) Probabilistic modeling of one-dimensional water movement and leaching from highway embankments containing secondary materials. Environ Eng Sci 22:156–169
45. Kosson DS, van der Sloot HA, Sanchez F, Garrabrants AC (2002) An integrated framework for evaluating leaching in waste management and utilization of secondary materials. Environ Eng Sci 19:159–204
46. Dabo D, Badreddine R, de Windt L, Drouadaine I (2009) Ten-year chemical evolution of leachate and municipal solid waste incineration bottom ash used in a test road site. J Hazard Mater 172:904–913
47. Reijnders L (2007) Cleaner phosphogypsum, coal combustion ashes and waste incineration ashes for application in building materials: a review. Build Environ 42:1036–1042
48. Aguiar del Toro M, Calmano W, Ecke H (2009) Wet extraction of heavy metals and chloride from MSWI and straw combustion ashes. Waste Manage 29:2494–2499
49. Sakanakura H (2007) Formation and durability of dithiocarbamic metals in stabilized air pollution control residue from municipal solid waste incineration and melting processes. Environ Sci Technol 41:1717–1722
50. Bosshard RP, Bachofen R, Brandl H (1996) Metal leaching from fly ash from municipal waste incineration by *Aspergillus niger*. Environ Sci Technol 30:3066–3071
51. Yang J, Wang QH, Wang Q, Wu TJ (2008) Comparison of one-step and two-step bioleaching for heavy metal removal from municipal solid waste incineration fly ash. Environ Eng Sci 25:783–789
52. Bayuseno A, Schmahl WW, Müllejans T (2009) Hydrothermal processing of MSWI fly ash – towards new stable minerals and fixation of heavy metals. J Hazard Mater 167:250–259
53. Fraissler G, Jollet M, Mattenberger H, Brunner T, Obernberger I (2009) Thermodynamic equilibrium calculations concerning the removal of heavy metals from sewage sludge ash by chlorination. Chem Eng Process 48:152–164
54. Bo D, Zhang F, Zhao L (2009) Influence of supercritical water treatment on heavy metals in medical waste incineration ash. J Hazard Mater 170:66–71
55. Guo X, Xiang D, Duan G, Mou P (2010) A review of mechanochemistry applications in waste management. Waste Manage 30:4–10
56. Siddique R (2008) Waste materials and by-products in concrete. Springer, London
57. Yao J, Li W, Kong Q, Wu Y, He R, Shen D (2010) Content, mobility and transfer behavior of heavy metals in MSWI bottom ash in Zhejang province, China. Fuel 89:616–622
58. Huang C, Yang W, Ma H, Song Y (2006) The potential of recycling and reusing municipal solid waste incinerator ash in Taiwan. Waste Manage 26:979–987
59. Pan JR, Huang C, Kao J, Lin S (2008) Recycling MSWI bottom and fly ash as raw materials in Portland cement. Waste Manage 28:1113–1118
60. Ferreira C, Ribeiro A, Ottosen L (2003) Possible applications for municipal solid waste fly ash. J Hazard Mater B 96:201–216
61. Toller S, Kärrman E, Gustafsson JP, Magnusson Y (2009) Environmental assessment of incinerator residue utilization. Waste Manage 29:2071–2077
62. Francois D, Pierson K (2009) Environmental assessment of a road site built with MSWI residue. Sci Total Environ 407:5945–5960
63. Huang W, Tang H, Lin K, Liao M (2010) An emerging pollutant contributing to cytotoxicity of MSWI ash wastes: strontium. J Hazard Mater 173:597–604
64. Dubey B, Townsend T (2007) Leaching of milled asphalt pavement amended with waste-to-energy ash. Int J Environ Waste Manage 1:145–158
65. Kayhanian M, Vichare A, Green PG, Harvey J (2009) Leachability of dissolved chromium in asphalt and concrete surfacing materials. J Environ Manage 90:3574–3580
66. Birgisdottir H, Bhander G, Hauschild MZ, Christensen TH (2007) Life cycle assessment of disposal of residues from municipal solid waste incineration: recycling bottom ash in road construction or landfilling in Denmark evaluated by the ROAD–RES model. Waste Manage 27:S75–S84
67. Tervahattu H, Kupiainen KJ, Räisänen M, Mäkelä T, Hillamao K (2006) Generation of urban road dust from anti-skid and asphalt concrete aggregates. J Hazard Mater 132:39–46
68. Travat I, Lidelow S, Anderas L, Tham C, Lagerkvist A (2009) Assessing the environmental impact of ashes used as landfill cover construction. Waste Manage 29:1336–1246
69. Beyer C, Konrad W, Rügner H, Bauer S, Liedl R, Grathwohl P (2009) Model-based prediction of long term leaching of contaminants from secondary materials in road constructions and noise protection dams. Waste Manage 29:839–850
70. Reijnders L (2007) The cement industry as a scavenger in industrial ecology and the management of hazardous substances. J Ind Ecol 11(1):15–25
71. Karstensen KH, Kinh NK, Thang LB, Vet PH, Tuan ND, Toi DT, Hung NH, Quan TM, Hanh LB, Thang DH (2006) Environmentally sound destruction of obsolete pesticides in developing countries using cement kilns. Environ Sci Policy 9:577–586

72. Sidhu S, Kast N, Edwards P, Dellinger B (2001) Hazardous air pollutants formation from reactions of raw meal organics in cement kilns. Chemosphere 42:499–506
73. Chen C (2004) The emission inventory of PCDD/PCDF in Taiwan. Chemosphere 54:1413–1420
74. Dermatas CM (2006) Evaluation of ettringite and hydrolumite formation for heavy metal immobilization. J Hazard Mater 136:20–33
75. van der Sloot HA, Seignette P, van Zomeren A, Hoede D, Meeuwsen JCL (2003) Effects of alternative materials, life cycle stages, testing and criteria development. www.ecn.nl
76. Winder C, Carmody M (2002) The dermal toxicity of cement. Toxicol Ind Health 18:321–331
77. Liden C (2001) Legislative and preventive measures related to contact dermatitis. Contact Dermat 44:65–69
78. Lannoye PA (2003) Report on proposed Directive of the European Parliament and Council regarding the limitation of marketing nonylphenol, nonylphenol ethoxylate and cement. European Parliament, Brussels
79. Costa M, Klein CB (2006) Toxicity and carcinogenicity of chromium compounds in humans. Crit Rev Toxicol 36: 155–163
80. Guo Q (1997) Increases of lead and chromium in drinking water from using cement–mortar–lined pipes: initial modeling and assessment. J Hazard Mater 56:181–213
81. Pedersen AJ, Frandsen FJ, Riber C, Astrup T, Thomsen SN, Lundtorp K, Mortensen LF (2009) A full-scale study on the partitioning of trace elements in municipal solid waste incineration – effects of firing different waste types. Energy Fuels 23:3475–3489
82. Alba N, Vazquez E, Gasso S, Baldasano JM (2001) Stabilization/solidification of MSW incineration residues from facilities with different air pollution systems. Durability of matrices versus carbonation. Waste Manage 21:313–324
83. Yvon J, Antenucci D, Lorenzi G, Dutre V, Leclerq D, Nielsen P, Veschkens M (2006) Long term stability in landfills of municipal solid waste incineration fly ashes solidified/stabilized by hydraulic binders. J Geochem Explor 90:143–155
84. Meima JA, Comans RNJ (1998) Reducing Sb leaching from municipal solid waste incineration bottom ash by addition of sorbent materials. J Geochem Explor 62:299–304
85. Valls S, Vazquez E (2001) Accelerated carbonation of sewage sludge–cement–sand mortars and its environmental impact. Cem Concr Res 31:1271–1276
86. Ecke H (2003) Sequestration of metals in carbonated municipal solid waste incineration (MSWI) fly ash. Waste Manage 23:631–640
87. Garrabrants AC, Sanchez F, Kosson DS (2004) Changes in constituent equilibrium leaching and pore water characteristics of a Portland cement mortar as a result of carbonation. Waste Manage 24:19–36
88. Idachaba MA, Nyavor K, Egiebor NO (2003) Microbial stability evaluation of cement based waste forms at different waste to cement ratios. J Hazard Mater B 96:331–340
89. Brombacher C, Bachofen R, Brandl H (1997) Biohydrological processing of solids. A patent review. Appl Microbiol Biotechnol 48:577–587
90. Yang J, Wang Q, Luo Q, Wang Q, Wu T (2009) Biosorption behavior of heavy metals in bioleaching process of MSWI fly ash by *Aspergillus niger*. Biochem Eng J 46:294–299
91. van Zomeren A, Comans RNJ (2009) Carbon speciation in municipal solid waste incinerator (MSWI) bottom ash in relation to facilitated metal leaching. Waste Manage 29:2059–2064
92. Twardowska J, Szczezepanska J (2002) Solid waste: terminological and long term environmental risk management problems exemplified in a power plant fly ash study. Sci Total Environ 285:28–51
93. Bayard R, Pestre C, Gourdon R (2009) Aerobic microbial activity in fresh and aged bottom ashes from municipal waste incineration. Int Biodeterior Biodegradation 63: 739–746
94. Aberg A, Kumpiene J, Ecke H (2006) Evaluation and prediction of emissions from a road built with bottom ash from municipal solid waste incinerator (MSWI). Sci Total Environ 355:1–12
95. van der Sloot HA (2000) Comparison of the characteristic leaching behavior of cements using standard EN–196–1 cement mortar and an assessment of their long-term environmental behavior in construction products during service life and recycling. Cem Concr Res 30:1079–1096
96. Serclérat J, Moskowicz P, Pollet B (2000) Retention mechanisms in mortars of trace metals contained in cement clinkers. Waste Manage 20:259–264
97. Hunsinger H, Seifert H, Jay K (2006) An economic process to inhibit PCDD/PCDF formation in MSWI by SO_2. Organohalogen Compd 68:151–156
98. Hunsinger H, Seifert H, Jay K (2007) Reduction of PCDD/F formation in MSWI by a process-integrated SO_2 cycle. Environ Eng Sci 24:1145–1159
99. Hunsinger H, Seifert H, Jay K (2007) Control of PCDD/F formation under conditions of fluctuating combustion performance in MSWI. Organohalogen Compd 69:956–961
100. Ke S, Ianhua Y, Xiaodong L, Shenyong L, Yinglei W, Muxing F (2010) Inhibition of *de novo synthesis of PCDD/Fs by* SO_2 in a model system. Chemosphere. doi:10.1016/j.chemosphere.2009.12.043
101. Mast P (1999) Einfluss der Abfallzusammensetzung auf Schadstofgehalt und– Menge der Verbrennuingsrückstaände (Impact of waste composition on the concentration and amount of toxics in combustion residues). TAUW, Berlin
102. Jeong SM, Osako N, Kim Y (2005) Utilizing a database to interpret leaching characteristics of lead from bottom ashes of municipal waste incinerators. Waste Manage 23:694–701
103. Lo S, Tsao Y (1997) Economic analysis of waste minimization for electroplating plants. Water Sci Technol 36:383–390
104. Fujimori T, Takaoka M, Takea N (2009) Influence of Cu, Fe, Pb and Zn chlorides and oxides on formation of chlorinated aromatic compounds in MSWI fly ash. Environ Sci Technol 43:8053–8059

105. Franz M (2008) Phosphate fertilizers from sewage sludge ash (SSA). Waste Manage 28:1809–1818
106. Adam C, Peplinski B, Michaelis M, Kley G, Simon D (2009) Thermochemical treatment of sewage sludge ashes for phosphorous recovery. Waste Manage 29:1122–1128
107. Wzorek Z, Jodko M, Gorazda K, Rzepecki T (2006) Extraction of phosphorus compounds from ashes from thermal processing of sewage sludge. J Loss Prev Process Ind 19:39–50
108. Donatello S, Freeman Pask A, Tyrer M, Cheeseman CR (2010) Effect of milling and acid washing on the pozzolanic activity of incinerator sewage sludge ash. Cement Concr Compos 32:54–61
109. Mattenberger H, Fraissler G, Herk P, Hermann L, Obernberger I (2008) Sewage sludge ash to phosphorus fertilizer: variables influencing heavy metal removal during thermochemical treatment. Waste Manage 28:2709–2722
110. Ottosen LM, Perdersen AJ, Hansen HK, Ribeiro AB (2007) Screening the possibility for removing cadmium and other heavy metals from wastewater sludges and bioashes by an electrodialytic method. Electrochim Acta 52:3420–2426
111. Christen C (2007) Closing the phosphorus loop. Environ Sci Technol 46:2078
112. Zhang F, Yamasaki S, Nazyo M (2001) Application of waste ashes to agricultural land – effect of incineration temperature on chemical characteristics. Sci Total Environ 264: 205–214
113. Lienert J, Larsen TA (2010) High acceptance of urine source separation in seven European countries: a review. Environ Sci Technol 44:556–566
114. Zhang D, Yamaski S, Kimura K (2001) Rare earth element content in various waste ashes and the potential risk to Japanese soils. Environ Int 27:393–398
115. Rosen CJ, Bierman PM, Olson D (1994) Swiss chard and alfalfa responses in soils amended with municipal waste incinerator ash: growth and elemental composition. J Agric Food Chem 42:1361–1368
116. Ferreira C, Ribeiro A, Ottosen L (2003) Possible applications for municipal solid waste fly ash. J Hazard Mater B 93:201–216
117. Pasquini MW (2006) The use of town refuse ash in urban agriculture around Jos, Nigeria: health and environmental risks. Sci Total Environ 354:43–59
118. Passquini MW, Alexander MJ (2004) Chemical properties of urban waste ash produced by open burning on the Jos Plateau: implications for agriculture. Sci Total Environ 319:325–340
119. Hwa TJ, Joyseelan S (1977) Conditioning of oily sludges with municipal solid wastes incineration fly ash. Water Sci Technol 35:231–238
120. Yue Q, Han S, Yue M, Gao B, Li Q, Yu H, Zhao Y, Qi Y (2009) The performance of biological anaerobic filters packed with sludge–fly ash ceramic particles (SPCP) and commercial ceramic particles (CCP) Turing the restart period. Effect of the C/N ratios and filter media. Bioresour Technol 100:5016–5020
121. Han S, Yue Q, Yue M, Gao B, Li Q, Yu H, Zhao Y, Qi Y (2009) The characteristics and application of sludge–fly ash ceramic particles (SCP) as novel filter media. J Hazard Mater 171: 809–814
122. Pan S, Lin C, Tseng D (2003) Reusing sewage sludge as absorbent for copper removal from wastewater. Resour Conserv Recycl 39:79–90
123. Bouzid J, Elouear Z, Ksibi M, Feki M, Montiel A (2008) A study on removal characteristics of copper from aqueous solution by sewage sludge and pomace ashes. J Hazard Mater 152:838–845
124. Okada K, OnoY KY, Nakajima A, MacKenzie KJD (2007) Simultaneous uptake of ammonium and phosphate ions by compounds prepared from paper sludge ash. J Hazard Mater 141:622–629
125. Wajama T, Haga M, Kuzawa K, Ishimoto H, Tamada O, Ito K, Nishiyama T, Downs RT, Rakovan JF (2006) Zeolite synthesis from paper sludge ash at low temperature (90°C) with addition of diatomite. J Hazard Mater B 132:244–252
126. Yang GCC, Yang T (1998) Synthesis of zeolites from municipal incinerator fly ash. J Hazard Mater 62:75–89
127. Shim Y, Kim Y, Kong S, Rhee S, Lee W (2003) The adsorption characteristics of heavy metals by various particle sizes of MSWI bottom ash. Waste Manage 23:851–857
128. Jin J, Chi L, Yan J (2010) Co-disposal of heavy metals containing waste water and medical waste incinerator fly ash by hydrothermal process with addition of sodium carbonate: a case study on Cu(II) removal. Water Air Soil Pollut. doi:10.1007/s 11270-009-0207-5
129. Ahmaruzzaman M (2010) A review on the utilization of fly ash. Prog Energy Combust Sci. doi:10.1016/j.pecs.2009.11.003
130. Smith KM, Fowler GD, Pulket S, Graham NJD (2009) Sewage sludge-based adsorbents: a review of their production, properties and use in water treatment applications. Water Res 43:2569–2594
131. Baciocchi R, Polettini A, Pomi R, Prigiobbe V, von Zedwitz VN, Steinfeld A (2006) Sequestration by direct gas–solid carbonation of air pollution control (APC) residues. Energy Fuels 20:1933–1940
132. Pertl A, Mostbauer P, Obersteiner G (2010) Climate balance of biogas upgrading systems. Waste Manage 30:92–99
133. Ducom G, Radu-Tirnoveanu D, Pascual C, Benadda B, Germain P (166) Biogas-municipal solid waste incinerator bottom ash interactions: sulphur compounds removal. J Hazard Mater 166:1102–1108
134. Wang S, Wu H (2006) Environmental–benign utilization of fly ash as low-cost adsorbents. J Hazard Mater B 136:482–501
135. Karatza D, Lancia A, Musmarra D (1998) Fly ash capture of mercuric chloride vapors from exhaust combustion ash. Environ Sci Technol 32:3999–4004
136. Reijnders L, Huijbregts MAJ (2009) Biofuels for road transport. Springer, London
137. Capello C, Hellweg S, Hungerbühler K (2008) Environmental assessment of waste–solvent treatment options. J Ind Ecol 12:111–127

138. Björklund A, Finnveden G (2005) Recycling revisited – life cycle comparisons of global warming impact and total energy use of waste management strategies. Resour Conserv Recycl 44:309–317
139. Luteijn J (2009) No energy to waste. Thesis, Open University of the Netherlands, Heerlen
140. Vehlow J, Bergfeldt B, Hunsinger H (2006) PCDD/F and related compounds in solid residues from municipal solid waste incineration – a literature review. Waste Manage Res 24:404–420

Hazardous Waste Incinerator Emissions

Montse Mari, José L. Domingo, Martí Nadal,
Marta Schuhmacher
Laboratory of Toxicology and Environmental Health,
School of Medicine, IISPV, Universitat Rovira i Virgili,
Reus, Catalonia, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
Chemicals of Concern
A Review of Scientific Publications on HWI Emissions
Environmental and Human Biological Monitoring of Pollutants
A Case Study: The HWI of Constantí (Catalonia, Spain)
Future Directions
Bibliography

Glossary

Environmental monitoring Control of the environmental status of specific location by means of analyzing the concentration of pollutants in environmental matrices (air, soil, water, or sediments).

Hazardous waste (HW) Materials, usually in solid or liquid form, which are generally produced in industrial processes, and characterized by posing substantial or potential threats to public health or the environment. They must be properly treated and/or conditioned in special HW management plants.

Human health risk assessment Evaluation of the probability that the human exposure to certain substance(s) can derive in adverse health effects to individuals or populations. A differentiation between non-carcinogenic and cancer risks is generally performed.

Incineration Typical process to manage municipal solid waste or hazardous waste consisting on its thermal destruction. Organic matter contained in waste is completely destroyed by combustion in special ovens with the presence of oxygen in excess. Waste is converted into ash, flue gas, and heat. Therefore, incineration is also known as waste-to-energy or energetic valorization.

Polychlorinated dibenzo-*p*-dioxins and dibenzofurans (PCDD/Fs) Simply called "dioxins," PCDD/Fs form a group of chemical substances whose molecular structure presents two benzene rings joined by two oxygen bridges. They include a family of 210 congeners, which are differentiated according to their number and position of chlorine atoms. They were originally included in the Stockholm Convention treaty that aims the elimination/restriction of Persistent Organic Pollutants (POPs). They are characterized by being highly persistent, bioaccumulative and toxic substances, and with a long-range transport capacity. PCDD/Fs are unintentional by-products of combustion processes. Two great groups of sources have been traditionally identified: regulated and nonregulated/diffuse.

Definition of the Subject

This chapter is basically focused on the assessment of emissions from hazardous waste incinerators (HWIs), evaluating specially those pollutants with the greatest interest and concern due their notable toxicity, persistence, and bioaccumulation capacity: polychlorinated dibenzo-*p*-dioxins and dibenzofurans (PCDD/Fs) and heavy metals. The main properties and toxic effects of those chemicals are here summarized. A review of publications on emissions from HWIs indicates that in the last decade significant progresses have been achieved in emission reductions of these facilities as result of very rigorous regulations. The suitability of using environmental (e.g., soils, herbage, etc.) and human biological (e.g., blood, urine, etc.) monitors to evaluate the impact of the above pollutants is presented together with some results of a case study, a wide survey of a HWI located in Catalonia (Spain). The results indicate that when HWIs are properly designed using the best available techniques (BAT) as well as the best environmental practices (BEP), the emissions are not environmentally significant and, consequently, they do not mean additional risks for the human health.

Introduction

Hazardous waste (HW) is defined as a waste that is dangerous or potentially harmful to human health and/or the environment. HWs include industrial wastes such as organic solvents, waste oil, sludge, rubber, etc. All these wastes contain toxic substances. Therefore, their proper management is crucial to protect the environment and safety of individuals. Currently, there is no estimation about global generation of industrial wastes. However, the rapid increase in volume and types of hazardous waste as a result of the continuous economic growth, urbanization, and industrialization is becoming a burgeoning problem for national, regional, and local governments to ensure effective and sustainable management of waste [1].

Incineration using best available technologies (BAT) in Europe, or maximum achievable control technologies (MACT) in USA, is considered the preferred management option for certain wastes that cannot be recycled or reused, being prevention the first step in the waste management strategy. Combustion is used to destroy hazardous organic constituents and to reduce the volume of waste. However, waste incineration, and specifically HW, generates a great concern on the population because of the toxic substances contained in wastes or other formed during the combustion process, which can be released through the stack, causing potential adverse effects on the environment and the human health of the population living nearby. This chapter is focused on pollutants potentially emitted by incinerators, as well as with the associated environmental and human impacts.

Stack emissions of hazardous waste incinerators (HWIs) contain major contaminants, those at higher concentrations such as nitrogen oxides (NO_x), sulfur dioxide (SO_2), hydrogen chloride (HCl), hydrogen fluoride (HF), particulate matter (PM), and volatile organic compounds (VOCs), including chloroform ($CHCl_3$), carbon tetrachloride (CCl_4), and chlorobenzenes (CLPs). In addition, stack emissions also include minor contaminants, those present in lower concentrations, such as persistent organic pollutants (POPs) and metals. However, because of their characteristics of persistence, toxicity, and capacity of bioaccumulation, POPs and metals are those of major concern. Consequently, these will be considered here in detail.

Some POPs found in stack gas of HWIs are hexachlorobenzene (HCB), polycyclic aromatic hydrocarbons (PAHs), polychlorinated dibenzo-*p*-dioxins and dibenzofurans (PCDD/Fs), polychlorinated biphenyls (PCBs), polychlorinated diphenyl-ethers (PCDEs), and polychlorinated naphthalenes (PCNs) [2]. Among them, PCDD/Fs are the most studied since they were the first ones discovered in the flue gases and fly ash of municipal solid waste incinerators (MSWIs) [3]. The two main methods for controlling PCDD/Fs emissions from waste incinerators are stable combustion and the removal of emitted PCDD/Fs by air pollution control devices (APCDs). The parameters that affect the emissions of PCDD/Fs are those related with the wastes, such as the content of PCDD/Fs or their precursors, or the presence of chlorine in the feed waste, as well as other related to the combustion process, such as retention time, temperature, turbulence in the combustion chamber, and oxygen availability.

The exact mechanism of PCDD/Fs (“dioxins”) formation is unknown. However, it is recognized that dioxins generally form from the thermal breakdown of organic materials in the presence of transition metals and chlorinated compounds. Dioxin formation takes place from *de novo* synthesis, as the flue gas from the incineration process cools down from the initial 1,000°C down to about 250°C. Dioxin formation occurs only in the zone of the incinerator where temperature is low and residual particles (fly ash) are present. Fly ash acts as a catalytic surface during the formation process. At those conditions, the components for dioxin formation are carbon, small organic compounds, metal ions, and inorganic chloride [4]. Copper and iron are the metals better catalyzing this reaction [5]. Chlorine may be incorporated into dioxins through elemental Cl_2 or as HCl. As a consequence of this, regulatory efforts have been undertaken in incinerators to avoid PCDD/F formation and release. For example, incineration of HW with a content of more than 1% of halogenated organic substances, expressed as chlorine, has to accomplish with certain operational conditions in order to destroy as many organic pollutants as possible. Combustion control is based on dioxin destruction at high temperatures, typically over 1,000°C, and residence times of greater than 2 s. Based on remarkable advances in both combustion control and APCD technologies, many countries have established guidelines and regulations for the optimal operation of waste incinerators and for PCDD/Fs control [6].

With respect to metals, a number of studies have shown that elements like antimony (Sb), arsenic (As), cadmium (Cd), cobalt (Co), copper (Cu), chromium (Cr), lead (Pb), manganese (Mn), mercury (Hg), nickel (Ni), thallium (Tl), and vanadium (V) are emitted by industrial, medical, and municipal waste incinerators [7]. After combustion in modern HWIs, metals contained in HW are mainly collected in bottom and fly ash, being only small quantity of metals discharged from the stack as particulate or vapor [8]. Notwithstanding, atmospheric emission of these elements is a matter of concern. Heavy metals emissions from municipal and hazardous waste incinerators are controlled primarily through the use of particulate collection devices (electrostatic precipitators, fabric filters, and wet scrubbers) or acid gas control systems (dry injection, spray dryer absorption, and wet scrubbing). Most metallic compounds are in the vapor phase within the incineration system, where the temperature is around 1,000°C. The metallic compounds tend to condense as the flue gas is cooled and become adsorbed onto fine particulate matter (generally submicron in size, i.e., 0.2–0.7 μm). The major fraction of heavy toxic metals in the flue gas exists as fine particulates, and it is effectively controlled by properly sized electrostatic precipitators or fabric filters. On the other hand, a portion of the more volatile metals such as Hg and Cd may remain in the vapor phase, depending upon temperature conditions, and other technologies are needed to for metal removal from the stream. The additional control for vaporized toxic metals consists on spray dryer absorption systems or wet scrubbers [9].

With regard to the regulations, according to the EU legislation (Directive 2000/76/EC), the distinction between hazardous and nonhazardous waste is based principally on the properties of waste prior to incineration or co-incineration, but not on the differences in emissions. The same emission limit values should be applied to the incineration or co-incineration of hazardous and nonhazardous waste, but different techniques and monitoring measures upon reception of waste should be retained.

The EU has introduced measures to prevent or reduce air, water, and soil pollution caused by the incineration or co-incineration of waste, as well as the resulting risks to human health. These measures specifically require permissions for incineration and co-incineration plants, and emission limits for certain pollutants released to air or to water (Directive 2000/76/EC). This Directive establishes the strictest emission limit for PCDD/Fs, which is set in 0.1 ng International Toxic Equivalents (I-TEQ)/Nm^3. The emission of PCDD/Fs must be measured once or twice yearly under normal good operation conditions, although the periodicity is sometimes increased until obtaining 3-monthly (or even monthly) data. The emission limits established by the Directive of other pollutants such as dust and metals and are summarized in Table 1.

The United States Environmental Protection Agency (US EPA) proposes national emission standards for hazardous air pollutants (NESHAP) in hazardous waste combustors. These combustors include hazardous waste burning incinerators, cement kilns, lightweight aggregate kilns, industrial/commercial/institutional boilers and process heaters, and HCl

Hazardous Waste Incinerator Emissions. Table 1 Emission limits established by the EU Directive 2000/76/CE for some pollutants

Pollutant	Limit
Total dust (mg/m^3)	10
Cd + Tl (mg/m^3)	0.05
Hg (mg/m^3)	0.05
Sn + V + Sb + As + Pb + Cr + Co + Mn + Ni + Cu (mg/m^3)	0.05

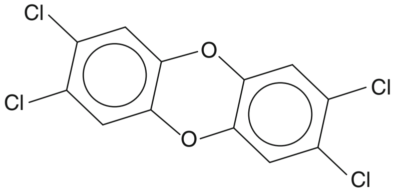
The results of the measurements shall be standardized at the following conditions: Temperature 273°K, pressure 101.3 kPa, 10% oxygen, dry gas production furnaces, known collectively as hazardous waste combustors (HWCs). The US EPA has identified these HWCs as major sources of hazardous air pollutant (HAP) emissions. These proposed standards will, when final, implement section 112(d) of the Clean Air Act (CAA) by requiring HW combustors to meet HAP emission standards reflecting the application of the maximum achievable control technology (MACT). US EPA regulations require sources to demonstrate a destruction and removal efficiency (DRE) for selected principal organic hazardous constituents (POHCs) of at least 99.99%; dioxin listed wastes are required to achieve at least 99.9999% DRE. Similar regulations are being developed in Korea [10], as well as in other countries.

Chemicals of Concern

Among the pollutants emitted by incinerators, dioxins and furans (PCDD/Fs) have traditionally generated the greatest concern because of their important toxicity, persistence, and bioaccumulation capacity. However, other POPs have been also found in HWI emissions. Polychlorinated biphenyls (PCBs) and polychlorinated naphthalenes (PCNs) are organochlorine compounds also formed in combustion processes, with similar characteristics to PCDD/Fs. In addition, incinerators can also emit other organochlorine compounds with toxic properties such as hexachlorobenzene (HCB) or chlorophenols (CLPs). A number of studies have shown that metals are also released in incinerator stacks. Among these, mercury is one of the most remarkable due to its high vapor pressure and high toxicity. A brief description of the sources of these compounds, as well as their behavior and toxicity are next described.

Organic Pollutants

Polychlorinated Dibenzo-*p*-Dioxins and Dibenzofurans (PCDD/Fs) The term polychlorinated dibenzo-*p*-dioxins and dibenzofurans (PCDDs), or simply dioxins, includes a set of 75 congeners, among which the most known and representative is the 2,3,7,8-tetrachlorodibenzo-*p*-dioxin or TCDD (Fig. 1). In turn, polychlorinated dibenzofurans (PCDFs), or simply furans, comprise 135 congeners. PCDDs and PCDFs (or PCDD/Fs) are two of the 12 persistent organic pollutants (POPs) (the “dirty dozen”) which were originally included in the Stockholm Convention treaty intended to reduce the environmental levels of POPs.



Hazardous Waste Incinerator Emissions. Figure 1 Chemical structure of TCDD

The 75 dioxin congeners mentioned above are due to the different position and/or the number of chlorine atoms in the aromatic rings. This number varies between 1 and 8. On the other hand, polychlorinated dibenzofurans or PCDFs have less symmetry positions, presenting a total of 135 congeners. It is precisely the number and position of chlorine atoms in the molecules of PCDD/Fs what determines the toxicity of the different compounds. Individual toxicity can vary with a relative value of 1,000, which obviously complicates the risk assessment of exposure, since they are commonly found as complex mixtures in the environment.

For this reason, and to simplify the risk evaluation, the concept of toxic equivalency factor (TEF) for each of the congeners that make up the mixtures was defined. The TEF concept, developed in the 1980s, provides a relatively simple method to evaluate the toxicity of complex mixtures of PCDD/Fs [11]. Several systems have been developed to express the total contribution of the 17 most toxic dioxins and furans. The toxicity of each congener is given relative to that of 2,3,7,8-TCDD (called Toxic Equivalency Factor, or TEF). Thus, to calculate the toxicity of the mixture, the concentration of each congener is multiplied by the corresponding TEF value of that congener. When these results are summed, a total TEQ or toxic equivalency of the mixture is obtained.

$$TEQ = \sum (TEF \times concentration)$$

The two most common systems to determine TEQs are based on the TEFs developed by the World Health Organization (WHO-TEF) [12] and the North Atlantic

Treaty Organization (NATO), known as International TEQ (I-TEF) [13]. Most congeners present similar toxicity equivalence factors in both systems, although some differences exist. Table 2 shows the different values of the I-TEFs and WHO-TEFs. The former are more common in North America, while Asia and Europe tends to use the latter.

Dioxins and furans may cause a wide range of biochemical and toxic effects in mammals. Apparently, the different congeners produce the same effects, but with varying degrees of potency. For most of them, the mechanism of action is not clear. However, it seems that the common denominator that mediates the biological effect of PCDD/Fs on cells is the binding of compounds to the aryl receptor hydrocarbon (AHR). Although carcinogenicity studies are still ongoing, one of the PCDD/F congeners, the 2,3,7,8-TCDD has been catalogued as probable human carcinogen by the US EPA, as some human studies have shown an association between this congener and soft tissue sarcomas, lymphomas, and stomach carcinomas. With regard to the non-carcinogenic effects, the most remarkable is the chloracne produced by the acute intoxication to dioxins. Nowadays, the toxicity of TCDD in humans is only partially known and mainly at short term (acute toxicity). The most representative signs and symptoms of acute and subacute exposure are the following:

Hazardous Waste Incinerator Emissions. Table 2 Values of I-TEF and WHO-TEF for the 17 most toxic PCDD/F congeners

Congener	I-TEF	WHO- TEF
2,3,7,8-TCDD	1	1
1,2,3,7,8-PeCDD	0.5	1
1,2,3,4,7,8-HxCDD	0.1	0.1
1,2,3,6,7,8-HxCDD	0.1	0.1
1,2,3,7,8,9-HxCDD	0.1	0.1
1,2,3,4,6,7,8-HpCDD	0.01	0.01
OCDD	0.001	0.0003
2,3,7,8-TCDF	0.1	0.1
1,2,3,7,8-PeCDF	0.05	0.03
2,3,4,7,8-PeCDF	0.5	0.3
1,2,3,4,7,8-HxCDF	0.1	0.1
1,2,3,6,7,8-HxCDF	0.1	0.1
1,2,3,7,8,9-HxCDF	0.1	0.1
2,3,4,6,7,8-HxCDF	0.1	0.1
1,2,3,4,6,7,8-HpCDF	0.01	0.01
1,2,3,4,7,8,9-HpCDF	0.01	0.01
OCDF	0.001	0.0003

Skin: chloracne, hyperkeratosis, and hyperpigmentation
Systemic: anorexia and weight loss, mild hepatic fibrosis, digestive problems, aching muscles and joints, pyelonephritis, cystitis hemorrhagic, etc.
Neurological and psychological: sexual dysfunction, visual problems, memory loss, loss of hearing, smell and touch, depression, irritability, sleep disorders
Others: eye irritation and conjunctivitis

Dioxins are not voluntarily produced, but they are formed as by-products in some chemical syntheses involving chlorine (i.e., paper making, manufacture of PVC or herbicides, etc.) and combustion processes (i.e., urban and hospital waste incineration, cement production, metal refining, etc.).

PCDD/Fs are highly persistent in the environment. By increasing the degree of chlorination, both the stability and the lipophilicity (solubility in oil) are enhanced. The octanol/water constant (K_{ow}), which indicates the relative lipid solubility in water, is in an order of one hundred million (log K_{ow} range: 6.5–8.8), which implies a high tendency to move to the fat. The soil microorganisms or animals are not able to effectively break the chlorine in the positions 2,3,7,8 of PCDD/Fs, which makes their elimination very slow.

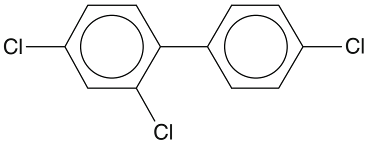
PCDD/Fs are highly persistent and lipophilic substances. They can be adsorbed very easily by particles present in air, soils, and sediments. There is an increase in its concentration along the food chain. Thus, they are present at highest concentrations in organisms of higher levels in the food chain (birds and mammals), ultimately accumulating in the human body. It is known that these compounds are biomagnifying. Moreover, since dioxins and furans are persistent and semi-volatile, they can be transported long distances by the wind, undergoing what is known as a Long-Range Atmospheric Transport (LRAT). As a consequence, it is sometimes possible to find PCDD/Fs in areas very distant from any known source of emission.

Polychlorinated Biphenyls

Polychlorinated biphenyls (PCBs) comprise a group of chlorinated aromatic hydrocarbons consisting of 209 congeners. As PCDD/Fs, the number and position of chlorine atoms determines the biological properties and environmental behavior of individual polychlorinated biphenyls (PCBs). As an example, the PCB-28 chemical structure is depicted in Fig. 2.

Commercial mixtures of PCBs have a wide spectrum of toxic effects. Many of these effects are very similar to those caused by 2,3,7,8-TCDD [14]. Some adverse health effects associated with PCBs are chloracne, changes in the immune system, as well as in neurobehavior of adults and children [15]. The United Nations Organization (UNO), through the WHO, has concluded that according to the degree of chlorination and the exposure time, PCBs produce adverse effects such as chloracne, increased eye discharge, swelling, and pain abs (Yushin disease) [16].

Studies in people occupationally exposed to PCBs show that they may cause lung and nasal irritation, as well as liver and hematologic and gastrointestinal disorders, depression, and fatigue [15]. Some PCBs alter thyroid hormone action and/or other endocrine glands that can affect growth, development, and reproduction. PCBs can also cause cancer in mammals [15]. Although it has not been proved that these chemicals cause cancer in humans, both the US EPA and the International Agency for Research on Cancer (IARC) have classified PCBs as probable human carcinogen (category 2A); that is to say, they are considered substances that have little evidence to have causal relationship with cancer in humans, but sufficient evidence presented in studies with laboratory animals [17]. Otherwise, some congeners have carcinogenic effects and may be associated with an increased risk of breast cancer [18].



Hazardous Waste Incinerator Emissions. Figure 2
Chemical structure of PCB-28

PCBs are not naturally found in the environment, but they are substances generated in technological processes. PCBs were synthesized in a laboratory for the very first time in 1929 in USA. Their manufacture produces a mixture of compounds, whose properties depend on the degree of chlorination. In general, they present high thermal, chemical, and biological stability with a high dielectric constant. They are insoluble in water and can mix with oils. They are also highly resistant to fire. According to those properties, they were widely used in the past in electrical equipment as insulators such as transformers and thermostats. In addition, they were also used for years in many industrial applications as the manufacture of adhesives, varnishes, plasticizers, and algaecides molusquicides.

Between 1929 and 1977, the total worldwide production of PCBs was around 1,200,000 t. Due to their flameproof characteristics, the majority of dielectric oils with PCBs were used in areas with high fire risk, such as petrochemical and industrial plants [19]. However, in 1977, the manufacture and use of PCBs was banned, as PCBs were noted to cause a major environmental impact because of the properties that made them so useful.

Due to its physical and chemical properties, PCBs tend to be adsorbed on the particles of air, soil, and sediment. Although the air is the compartment where they are degraded more easily, some studies have suggested lifetimes ranging from 2 to 6 years depending on the degree of chlorination [20]. In this compartment, the photodegradation is considered the most important mechanism of elimination [21]. In water, PCBs also adsorbs to particles where they can remain for years [22]. PCBs show a high affinity for soils and sediments that act as reservoirs of these compounds. High lifetimes are observed in this compartment especially for the heavier compounds [23].

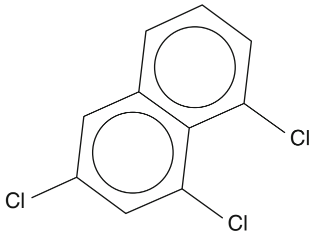
PCBs are highly hydrophobic compounds with high lipid solubility, so they tend to accumulate in the adipose tissue. Due to its ability to solubilize lipids and the high persistence and capacity for bioaccumulation, biomagnification occurs along the food chain, increasing their toxicological effects in organisms. Regarding the biological lifetimes, there is a great disagreement on the different congeners of PCBs. It is believed that the lifetimes of commercial mixtures of PCBs in the human body can range from 2 to 6 years [24]. However, if the lifetimes of different congeners are individually measured, these values can vary greatly. For example, the average life of PCB-153 is estimated in 12.4 years, whereas that of PCB-28 is estimated in 1.4 years [25].

Like PCDD/Fs, PCBs have a great tendency to be adsorbed on the particles of air. Thus, they tend to move up very long distances. As a result, concentrations of PCBs have been detected in air and other compartments in very remote areas like the Arctic mountains, far from any source [26].

Polychlorinated Naphthalenes

Polychlorinated naphthalenes (PCNs) form a family of 75 congeners whose chemical structure is composed by naphthalene with one to eight chlorine atoms. For example, Fig. 3 shows the chemical structure of trichloro-naphthalene. The physical and chemical properties of PCNs are very similar to those of PCBs. They are hydrophobic and present a high chemical and thermal stability. Moreover, they are good electrical insulators and flame retardants [27].

The scientific literature on the toxicity of PCNs is very limited. However, what is known is that, like other POPs, the mechanism of toxicity action is through the aryl hydrocarbon receptor (Ahr) [28]. According to human poisoning episodes, the effects of exposure to PCNs are chloracne, carcinogenesis, teratogenesis, and even mortality [29]. However, the fact that they always appear together with PCBs and PCDD/Fs may have caused an overestimation of the potential adverse effects of PCNs [28]. As in the case of PCDD/Fs and PCBs, PCNs are usually mixtures of different congeners. To characterize the toxicity of PCNs, TEF values have been established for some congeners. The TEFs for PCNs are 3–6 orders of magnitude less toxic than those of PCDD/Fs, while its potency is similar to that of PCBs [30]. Among them, some hexa-congeners are the most potent congeners, with a TEF of 0.002 [28]. Nevertheless, due to lack of sufficiently reliable data, the toxicity of PCNs has been quantified only for a few congeners [31, 32].



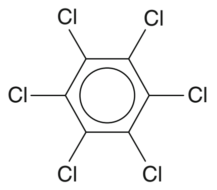
Hazardous Waste Incinerator Emissions. Figure 3
Chemical structure of trichloro-naphthalene

PCNs have been found in the gases and ashes from different incinerators in Sweden, Spain, and Germany [33–35]. In Canada, combustion processes involve the 54% of air releases of PCNs [19]. Chlor-alkali processes have been pointed out as important sources of PCNs [33, 36]. Other sources are technical PCB formulations [37].

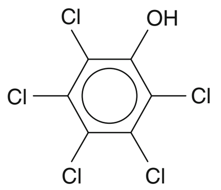
Historically, there have been several peaks of PCN wastes according to their degree of chlorination. The highest input of penta- and hexa-CNs occurred before 1944, while the peaks of tetra- and tri-CNS did not occur until the years 50 and 70, respectively [38]. However, a significant increase has been observed for many congeners associated with combustion, which suggests that, nowadays, combustion sources are more important than in the past [38].

The high octanol/water distribution constant (K_{ow}) indicates its capacity for biomagnification. The distribution constant octanol/water is especially significant for hexa- and hepta-CN homologues (log K_{ow} is between 3.90 and 8.3) [39]. Bioaccumulation experiments have focused mainly on determining the bioconcentration factors (BCF) in some aquatic species [40], being some BCF values higher than 20,000 [41]. It is believed that PCNs accumulate along the food chain and eventually accumulate in some human tissues and fluids [42]. Data on residence time of PCNs in water, soils, and sediments are very limited. In fact, lifetimes have been only established for mono-and di-CNS, with values from 38 to 104 days [43]. However, there is still much uncertainty around this issue.

During the preparation of the protocol of the European Convention on Long-Range Transboundary Air Pollution on POPs in the United Nations Economic Commission (UN-ECE-LRTAP POPs Protocol) 1998, the PCNs were not included into the list of the "Dirty Dozen" due to lack of appropriate data [44]. However, scientific knowledge obtained in recent years indicates that PCNs have become candidates to extend the list of chemicals to be restricted [40].



Hazardous Waste Incinerator Emissions. Figure 4 Chemical structure of chlorobenzene



Hazardous Waste Incinerator Emissions. Figure 5 Chemical structure of pentachlorophenol

Hexachlorobenzene

Hexachlorobenzene (HCB) is an organochlorine fungicide introduced for the first time in 1945 (Fig. 4). Many developing countries still use it to spray wheat. Currently, HCB is used as a plasticizer and as an additive for wood preservatives, being an important commodity for the synthesis of chlorinated organic compounds. HCB does not exist naturally in the environment, but it is obtained by chlorination of less chlorinated benzenes, being the basis for the production of pentachlorophenol. It can be also obtained as a by-product in the chlorinated hydrocarbon industry. Therefore, HCB enters in the environment by the burning of chlorine-containing products (e.g., incineration) or through the use of pesticides contaminated with this substance [45, 46].

Hexachlorobenzene is absorbed in the gastrointestinal tract, being slowly metabolized and accumulated in adipose tissue. When adipose tissue is mobilized, HCB is remobilized and then it can be found in all organs. In rats, prolonged exposure to this substance has produced, in some cases, skin diseases, neurological symptoms, and liver injury. In mice, immunotoxic effects were observed as well as viral infections in lungs and liver [47]. The Department of Health and Human Services (DHHS, USA) has determined that HCB could be potentially carcinogenic, since in animal studies liver, kidney, and thyroid cancers have been observed. Therefore, there is strong evidence that it may cause cancer in humans, but mutagenic or teratogenic effects have not been observed.

Chlorophenols

Chlorophenols (CLPs) are a family of 19 different compounds, of which the pentachlorophenol (PCP) is the most important (Fig. 5). Due to the wide spectrum of antimicrobial properties of chlorophenols, they have been used to protect wood, paint, plant fibers, and leather and as disinfectants. They have also been applied as herbicides, fungicides, and insecticides, as well as intermediary agents in the manufacture of pharmaceuticals and dyes.

Most CLPs are obtained by direct chlorination of phenol using chlorine gas. The final product has other isomers of chlorophenol as impurities. Therefore, depending on the conditions of process, chlorophenols can generate heavier polychlorinated phenoxy phenols and low concentrations of PCDD/Fs, being possible the emissions in their manufacture, storage, transport, and application [48].

Chlorophenols can be absorbed by lungs, gastrointestinal tract, and skin. Approximately 80% is expelled by the kidneys without having undergone any transformation. Its toxicity depends on the degree of chlorination, the position of chlorine atoms, and the purity of the sample. At low doses, they can cause irritation in the eyes and airways. Toxic doses produce convulsions, coma, and eventually death. After repeated administration, toxic doses can affect internal organs and bone marrow. Pentachlorophenol exerts a toxic effect on embryos in experimental animals. As a result of studies conducted in animals, which show an increased cancer risk, especially in the liver, the IARC considered that the PCP is a possible human carcinogen.

Heavy Metals

Arsenic Arsenic (As) is an element that is usually classified among metal, although it presents metalloid and nonmetal properties. This element is naturally found in the earth's crust, with levels between 2 and 5 mg/kg [49]. However, anthropogenic sources are usually more important than the natural sources. One of the main routes of entry of this element in the soil first, and afterward in water is through the use of fertilizers and pesticides in agriculture [50]. The air release is mainly produced by the combustion of coal and timber, mining, smelters, and waste incineration. Arsenic can be found in both inorganic and organic forms. While inorganic forms – As_2O_3, As_2O_5, or AsH_3 – are more toxic, organic arsenic, much less toxic, is originated as a form of detoxification through biotransformation by different organisms [51].

The exposure to As occurs mainly through food intake, especially seafood [52]. Fortunately, the percentage of organic arsenic (usually in the form of arseniobetaina) in foodstuffs is significantly much higher than that of inorganic As [53].

Arsenic compounds have been recognized as carcinogenic to humans (Group 1) based on different evidences [54]. Epidemiological studies have revealed a relationship between high concentrations of As in water and high incidences of skin cancer [55]. It has been also found a positive correlation between inhalation of As and lung cancer. The development of liver cancer, kidney, and bladder has also been associated with exposure to this element [56].

Beryllium Beryllium (Be) is a metal found naturally in rocks, coal, soil, and volcanic dust. The general population is exposed to low levels of Be through air, food, and water. Exposure to high levels of Be (greater than 1,000 mg/m) can produce symptoms similar to pneumonia. In general, however, it is unlikely that the general population develop chronic Be disease because environmental levels of this metal are usually very low (0.00003–0.0002 mg/m^3) [57].

No effects have been reported after ingestion of beryllium in humans because very little Be is absorbed by the stomach and intestines. Contact with skin may cause ulcers [57], while Be has been identified by the US EPA [58] as a probable human carcinogen (Group 2).

Cadmium Cadmium (Cd) is not an essential element from a nutritional point of view. It is released as a byproduct in zinc, copper, or lead smelting. It has excellent corrosion resistance, and therefore, a large percentage of the amount consumed is used to protect other metals (galvanized). A considerable amount is used to prepare pigments for the manufacture of paints and plastics and as a stabilizer of the latter. The mechanical properties of certain copper alloys improve with the addition of cadmium, and its use is widespread in the manufacture of alkaline batteries [59]. The skin absorption is minimal, whereas gastrointestinal absorption is 5% of the amount ingested, reaching up to 20% in case of iron or calcium deficiency. The absorption through inhalation can reach 50% of the total exposure, depending on the size of particles and their solubility [60].

Diet and tobacco play an important role in Cd exposure [61]. Some plants, such as rice or wheat, selectively concentrate cadmium present in soil and water. Thus, environmental pollution can reach the humans through the consumption of contaminated food. One of the greatest episodes of mass poisoning by cadmium was Itai-Itai disease in Japan at the end of World War II, caused by the consumption of rice and contaminated water [62].

The kidney is the organ primarily involved in chronic exposure to cadmium. The effects of inhalation also include pulmonary emphysema, anemia, dysfunction of liver, as well as dysfunctions in calcium, phosphorus, and vitamin D metabolisms. These effects are associated with osteoporosis, osteomalacia, and renal calcium stone formation [63]. On the other hand, the liver is the target organ for cadmium to settle, altering its subclinical hepatic function, and often producing a slight elevation of serum liver enzymes [63].

Finally, numerous epidemiological studies have shown that individuals that have been exposed for long periods to Cd have a higher risk than the general population of contracting prostate, lung, or breast cancer [61]. This element has been classified by the IARC as carcinogenic to humans (Group 1) and probable human carcinogen by the US EPA (Group B1).

Cobalt Cobalt (Co) is naturally found in rocks, soils, water, plants, and animals. Cobalt alloys are used for the manufacture of aircraft engines, magnets, grinding and cutting tools, and artificial knee and hip joints. Co compounds are also used to color glass, ceramics, and paintings to dry enamels and paintings on porcelain. Usually, Co is released to environment from natural sources, when burning coal or oil, or during the production of cobalt alloys. Once in air, cobalt is associated with particles and is finally deposited in a few days.

Cobalt has beneficial (is part of vitamin B12) and harmful effects to human health. Exposure to high levels of cobalt can have effects on lungs and heart [64], as well as it can also cause dermatitis. In animals exposed to high levels of cobalt, effects on liver and lungs have been reported [65]. It has been observed that non-radioactive cobalt does not produce cancer in humans or animals after exposure through food or water. However, cancer has been observed in animals after exposure through inhalation at high levels. Based on laboratory data from the IARC, it has been determined that Co and its compounds are possible carcinogens in humans.

Copper Copper (Cu) is a metal found naturally in rocks, soils, water, and air. Copper is an essential element for animals and plants, and therefore, necessary for life. It is used to manufacture many products such as wire or taps. It is also used in agriculture to treat plant diseases. Once released into the environment, Cu particles adhere to soil organic matter.

High levels of Cu can be harmful to health. Inhalation of high concentrations may cause irritation of the nose and throat. The ingestion of high levels of copper can cause nausea, vomiting, and diarrhea. Very high quantities can damage the liver, kidneys and even cause death [66]. It is not known whether copper can cause cancer in humans. Therefore, the US EPA has determined that copper is not classifiable as carcinogenic in humans [66].

Chromium Chromium (Cr) is found on the surface of rocks, soils, and the sea. This element is widely used in the industry, being its use quite extensive. As metal, it is not basically toxic. However, its oxidation products (chromic acid, chrome dicromats) are very toxic due to the caustic action. The major toxic effects are directly linked to the presence of trivalent chromium and, especially, hexavalent (more toxic) chromium, which are denaturing agents of proteins and nucleic acids, which are slowly eliminated into the urine [67, 68].

The cutaneous manifestations of Cr toxicity consist on skin ulcers and allergic dermatitis. Inflammation of the airways can also occur by inhalation [69]. Based on studies showing sufficient evidence that exposure to Cr (VI) causes lung cancer, the Cr (VI) has been classified by the IARC [70] as carcinogenic to humans (Group 1), while the US EPA has classified the hexavalent form of Cr as a known human carcinogen. Chronic exposure to Cr (VI) can produce non-carcinogenic effects in liver and kidney, as well as in the gastrointestinal and immune systems.

Tin Tin (Sn) metal is not soluble in water. It is present in brass, bronze, and some materials for welding. Tin is used to coat metal cans of food, drinks, and aerosols. Inorganic tin compounds are also used in toothpastes, perfumes, soaps, food additives, and dyes. Organic tin compounds are employed to manufacture plastic packaging for food, plastic pipes, pesticides, paint, and animal repellents. Tin is released into the environment by natural processes and human activities such as mining, oil and coal combustion, and the production and use of tin compounds. In the atmosphere, Sn can exist in vapor form, adhering to dust particles and soils. Organic tin compounds are bioaccumulative, as they can accumulate in animals (including fish) and plants.

Tin metal does not present a high toxicity because only a little part is absorbed from the gastrointestinal tract. Studies in humans have shown that ingestion of large amounts of inorganic tin can cause stomach pain, anemia and alterations in liver and kidneys. Inhalation of organic tin compounds may cause neurotoxic effects. In severe cases, it can cause death [71]. Some organic tin compounds affect the immune system in animals, but their effects in humans have not been accurately evaluated. Animal studies have also shown that tin compounds affect reproduction, although the evidence in humans is not clear. Contact with skin or eyes of organic or inorganic tin compounds can cause irritation [71]. There is no evidence that tin or its compounds may cause cancer in humans. Studies of inorganic compounds in animals have been negative. With respect to this, the IARC and the US EPA have not classified tin metal or its inorganic compounds as carcinogenic [58].

Manganese Manganese (Mn) is an element found naturally in the environment, usually in form of oxides. It is found in soil, air, water, and food. Anthropogenic sources of Mn include cement production plants, power plants, waste incineration, and combustion of fossil fuels [72]. In humans, Mn is an essential element.

Dietary intake is the main route of nonoccupational exposed individuals. Occupational exposure occurs mainly by inhalation of dust. Chronic inhalation of Mn causes effects in the nervous system [73]. There is no data confirming that the Mn causes cancer in humans. Thus, the US EPA does not classify manganese as a carcinogen metal [58].

Mercury Sources of mercury (Hg) are both natural and anthropogenic. Mercury is a liquid metal at room temperature, being the only metal that remains liquid at 0°C. Because of its high volatility and low boiling point, in the atmosphere it is in the gas phase [74]. In combination with elements such as oxygen, sulfur, or chlorine, it forms inorganic compounds. Mercury can be also combined with carbon to form organic mercury compounds. The most common of these is methyl mercury, which is produced by microorganisms in soil and water [75].

Medical and municipal waste incinerators, cement factories, and chlor-alkali plants are causing major concentrations of mercury vapor into the atmosphere [74]. Many occupational exposures occur by inhalation of mercury vapors. Chronic exposures to low concentrations of mercury are common [76].

The most important case of poisoning by methyl mercury was produced by ingestion of contaminated fish in the Minamata Bay, Japan. This episode was so important that the clinic profile caused by this intoxication was named Minamata disease [77]. The effects of mercury are highly dependent on their different chemical forms. The central nervous system is very susceptible to all forms of mercury. Methyl mercury and vapors of metallic mercury are more harmful than other forms because a greater amount of the former can reach the brain. Exposure to high levels of metallic, inorganic, or organic Hg can cause damage to the kidneys, brain, and fetus [75]. The neurological effects of mercury poisoning include involuntary shaking of limbs, lips, and tongue, and psychological disorders such as nervousness, irritability, abnormal reserved behavior (eretisme), mental confusion, and peripheral polyneuropathy, affecting the upper extremities more than the lower [78]. The US EPA has classified both inorganic mercury and methyl mercury within Group C (possible human carcinogens), while metallic mercury is not considered as a human carcinogen (Group D).

Nickel The main uses of nickel (Ni) are industrial. These include the manufacture of steel and other alloys and superalloys. It is used in smaller amounts for the manufacture of cast iron, chemicals, batteries, industrial ceramics, and catalyst. It is also considered an essential element for humans, although the specific deficiency has not been characterized yet.

Most occupational exposure to Ni occurs through inhalation of smoke or dust, or through dermal contact with the skin alloys, salts, or solutions [79]. Among workers of Ni-refining industries, a higher incidence of lung cancer and nasal cavities has been reported. Some studies indicate that these workers may suffer higher risk of getting other types of malignant processes such as cancer of the larynx, stomach, prostate, kidney, and soft tissue sarcoma.

Lead Because of its malleability and low melting point, humans have used lead (Pb) from prehistory in jewelry, manufacture of pipes, and water jars for drinking. Likewise, its salts are appreciated for their bright colors. Currently, it is widely used in industry to manufacture pigments, coatings, containers, ointments, and electric batteries.

The toxic effects of Pb are the same either if it is ingested or inhalated. Lead can affect almost all organs and body systems. In adults, prolonged exposure can impair some functions of the central nervous system [80]. In children, chronic exposure is associated with behavioral problems, learning difficulties and hyperactivity [81]. It can also cause weakness in fingers, wrists, and ankles. Exposure to lead produces an increase in blood pressure and can cause anemia [82] and damage to the kidneys [83]. Lead was classified as a possible carcinogenic in humans (Group 2B) by the IARC [54].

Thallium Thallium (Tl) is a metal found in small quantities in the Earth's crust. It is mainly used for the manufacture of electronic switches and terminals mainly in the semiconductor industry. It is also used in the manufacture of special glass and certain medical procedures. Thallium enters the environment mainly from burning coal and foundries in both cases as a contaminant of raw material. In the environment, Tl bioaccumulates, being absorbed by plants and entering the food chain.

Exposure to high levels of Tl can produce adverse effects on health. In a study in workers exposed for several years to inhalation of thallium, toxic effects were observed in the nervous system [84]. In studies of subjects who have ingested large amounts of thallium over a short period, vomiting, diarrhea, temporary hair loss, and neurological effects, and effects on lungs, heart, liver, and kidneys have been reported [84]. Acute exposure to high amounts of Tl also can cause death. Ingestion of thallium during long periods of time is not well known. Experimental studies with rats exposed to high levels of thallium have indicated developmental alterations. No information is available, either in humans or animals, on the Tl effects in contact with the skin [85]. There are no studies in humans or animals on the carcinogenic effects of inhalation, ingestion, or dermal contact with thallium. Neither the IARC nor the US EPA have classified Tl in terms of carcinogenesis in humans [58].

Vanadium Vanadium (V) and its compounds are present in the earth's crust and rocks, and it is recovered as a by-product in mining of other metals. The main use of this element is in alloys with iron, aluminum, and molybdenum, and to produce bronze and brass. Modest quantities of vanadium are also used as catalysts to make pottery, dyeing processes, insecticides, digital photography, color televisions, and glass production. Oil is relatively rich in this element, and the burning of large masses represents a considerable environmental contribution. Inhalation of dust or fumes, particularly as pentoxides, is the most common form of occupational exposure [86].

Vanadium is mainly absorbed through the respiratory tract and distributed to all tissues through the blood. It is found mainly in bone tissue, as metavanadate. It also binds to liver and lung, being slowly eliminated. The largest amount of vanadium is excreted by the kidney. The classic pattern of response to dust exposure at the working place with high V contents is the irritation of the upper respiratory tract. Depending on the V concentration in the air, symptoms appear within hours of exposure. The specific response includes coughing, forced breathing, eye irritation, nose and throat, bronchitis, chemical pneumonitis, and rarely hemorrhagic. Vanadium has not been implicated in carcinogenic processes [58].

A Review of Scientific Publications on HWI Emissions

Data reported in the scientific literature on emissions from HWIs are rather limited. This section is focused only on the publications from the last decade. The reason is that in the last years significant progresses have been achieved in emission reductions of waste incinerators, as a result of stringent regulations. Most papers discuss measures in combustion control and the performance of different air pollution control devices for POPs and/or metals. Furthermore, advances in the methodologies for stack emissions measurements of PCDD/Fs have been recently reported. Hinshaw and Trenholm [87] compiled data regarding stack emissions from various research and compliance testing programs from HWIs. As expected, it was found that the values for most pollutants considerably vary from incinerator to incinerator, according to operating conditions and feed composition. In relation to the main toxic metals (present in the ppb range), they included lead, chromium, cadmium, and arsenic. Mercury was also frequently found in this range, even though it was present at higher concentrations when emitted by other types of incinerators. Beryllium, if present, was found at lower concentrations in most cases. Naphthalene and other PAHs, as well as phenols, were found generally at higher levels (range 0–10 parts per billion – ppb). With respect to PCDD/Fs, their relative concentrations were the lowest in comparison with all the rest of pollutants. However, PCDD/Fs are especially toxic and bioaccumulative, and therefore one of major concern. van Velzen et al. [88] published a review on mercury and waste incineration, where the efficiency of different processes to remove mercury from flue gases was discussed (i.e., electrostatic precipitators, wet scrubbers, wet scrubbers with conditioning agent, spray absorbers, fabric filters, entrained flow adsorbers, and circulating fluidized bed). In that revision, examples of hazardous waste incinerators achieving good efficiencies were reported through the combination of diverse equipments. In all cases, the mercury outlet concentration was clearly below the limit of 50 μg/Nm^3. Löthgren and van Bavel [89] studied the dynamic behavior of the scrubber installed in a Swedish HWI since high concentrations of PCDD/Fs were found after the scrubber, even though the concentrations before the scrubber were low. They concluded that the high PCDD/Fs emissions after the installation of the wet scrubber were caused by memory effects due to high formation in the incinerator during the start-up of the plant. These periods with generation of high levels of dioxins during the start-up are not normally recognized since the sampling of PCDD/Fs normally takes place only during stable operating conditions. Consequently, these authors suggested the continuous monitoring of PCDD/Fs in order to cover all modes of operation, including those with elevated dioxin emissions. Gullet et al. [5] had previously emphasized the importance of previous firing conditions on the emissions, resulting in uncertainty regarding the prediction of emissions based solely on the type and rate of co-fired HW. Recently, Choi et al. [6] investigated the PCDD/F concentrations in MSWIs and HWIs from Korea. Different types of wastes and different APCD combinations were evaluated. The range of PCDD/F concentrations was, in general terms, much higher than the regulated upper limit (0.1 ng TEQ/Nm^3). The results indicated that the concentrations of PCDD/Fs and dioxin-like PCBs from small incinerators were notably higher than those from large incinerators. They concluded that a number of incinerators were operating in ways not intended in their original design (for instance, waste composition, shorter operating time, lower combustion temperature, broken APCDs), what was determinant in their emissions. Kim et al. [10] reported the current management practices and emission status from Korean HWIs in order to develop MACT-like guidelines. They studied 3 HWIs, which burned different industrial wastes. The measured air pollutants included dust, heavy metals (Be, Cd, Cr, As, Pb, Ni, Zn, and Hg and its speciation), and VOCs. All the tested HWIs were within the existing Korean limits. The main heavy metals emitted were Ni, Cr, Zn, and Pb. Metal composition in emissions of each one of the HWIs was mainly related to the type of wastes burned, but also to the operating parameters and type of APCDs. Metal removal efficiency of existing APCD was above 99%, excepting mercury for which it was above 30%. All metals sampled were removed within the particulate control devices, with the exception of mercury, which exhibited a higher concentration in the flue gas. Due to its lower boiling point this element escapes the particulate control devices. Benzene, dichloromethane, styrene, ethylbenzene, toluene, *m,p*-xylene, and *o*-xylene were among the VOCs emitted from the stacks of the tested HWIs. The VOC concentrations, as single compounds were low, but in aggregate their contribution was higher. Recently, Van Caneghem et al. [90] compared the amounts of PCDD/Fs and other POPs in the input and the output of three BAT waste incinerators. They developed a methodology to weight the POPs in the input and output of waste incinerators based on minimal risk doses or cancer potency factors. The ratios of the weighed POP input over output clearly depended on the kind of incinerator. The mass balance for PCBs indicated that they were effectively destroyed during incineration in a rotary kiln. This kind of incineration was identified as a weighed POP sink. On the other hand, the grate furnace incinerating MSW was a weighed POP sink or source depending on the POP concentrations in the waste. However, the difference between the output and the input was rather limited. Previously, Kim et al. [91] found that the mean concentration of PCDD/Fs in the stack gas of a HWI in Korea was 1.91 ng I-TEQ/Nm^3. They considered that HWIs were probably the major source of PCDD/F emissions in Korea because of their high emission factors which was due to the poor operation and inadequate maintenance. The congener profiles were characterized by relatively high levels of OCDD, OCDF, 1,2,3,4,6,7,8-HpCDF, and 1,2,3,4,6,7,8-HpCDD. In a recent study, Rivera-Austrui et al. [2] monitored the emissions of a HWI in Spain using a mew methodology that allows measuring PCDD/Fs emitted by the stack in a continued way, instead of punctually. So far, the common practice, which is set by the Directive 200/76/EC, is sampling during periods of 8 h. The new continuous measurement was thought to be a more representative sampling, allowing a more precise assessment of the real emissions and potential impact on the environment. The continuous sampling was carried out on a monthly basis with a total number of 16 samples covering a period of 2 year emissions. In addition, short-term samples were collected in order to compare them with the continuous results. The monthly results varied from 0.11 to 0.077 ng TEQ/Nm^3, resulting in an average of 0.30 ng TEQ/Nm^3. With respect to the short-term sampling performed during the same period of time, the results varied from 0.007 to 0.021 ng TEQ/Nm^3 (mean 0.012 ng TEQ/Nm^3). The average pattern of both methods was very similar. Nonetheless, subtle differences in the dioxin pattern were identified. The general PCDD/F patterns showed a higher amount of PCDFs than PCDDs in the flue gas. It was concluded that the emission factor for PCDD/F from "state-of-the-art" HWIs is largely overestimated compared with data continuously measured during 2 years.

Environmental and Human Biological Monitoring of Pollutants

As indicated above, few studies have investigated source samples. In contrast, environmental and biological monitoring is often used to evaluate the impact of pollutant sources and the human health risks. Environmental monitoring programs are common tools to control the state of pollution in a particular area or zone [92, 93]. These programs are used to assess the contribution of a potential contaminant source on the surrounding environment. Pollutant levels are determined in environmental matrices like air, water, soil, food, etc. For example, in industrial hygiene, air sampling has become prominent to evaluate workers exposure. For environmental impact, soil and vegetation are two of the most standardized monitors, since they can be easily collected and stored [94–96]. The atmospheric concentrations of environmental pollutants vary according to a number of factors, including the prevailing meteorological conditions. Because emissions of metals and PCDD/Fs from HWIs result in subsequent aerial deposition onto soil and plants, most monitoring programs utilize samples of these monitors. Soil and vegetation have been widely used as matrices of long-term and short-term accumulation, respectively, to environmental pollutants [97, 98]. The combination of different monitors is commonly used to get complementary information regarding punctual emissions (air), emissions during short periods of time (herbage reflects the deposition of pollution during few months), and emissions during long periods of time (soil reflects deposition during years). Environmental monitoring is especially necessary to identify sources of exposure and to facilitate measures for minimizing emissions. On the other hand, human biological monitoring has become an important tool in environmental and occupational medicine for the assessment of the levels of internal exposure to harmful substances taken up from general and occupational environments. Human biological monitoring may be used not only to evaluate the degree of internal exposure to an environmental pollutant of individuals or population groups, but also to evaluate temporal changes in the general population potentially exposed to a defined environmental contaminant [99–101]. It considers all routes of uptake and sources which are relevant. Human biomonitoring allows estimating the dose that is really taken up from the environment (internal dose). The internal doses of the pollutants in the different body tissues depend on their biokinetics and metabolism. The levels of rapidly disappearing biomarkers primarily reflect exposures during the initial hours. Moreover, biomarkers which disappear over the course of several weeks reflect one, several, or numerous exposure incidents occurring anytime during a period of various weeks previous to the measurement. Some toxicants accumulate in one or several tissues and are in dynamic equilibrium with the sites of toxicity. In the case of PCDD/Fs which basically accumulate in fats, the plasma PCDD/F concentrations reflect the amount stored in the body.

According to Angerer et al. [102], the requirements for a proper human biomonitoring of a certain chemical are:

(a) Suitable biomonitoring matrices
(b) Suitable parameters, which can be able to reflect internal exposure
(c) Suitable and reliable analytical methods which are kept under control by quality assurance
(d) Reference and limit values which enable the interpretation of results

Because accuracy is crucial, as trace levels are involved, to determine concentrations of chemicals and metabolites in biological samples require meticulous attention in both sampling process and analytical techniques. Biological monitors include not only blood and urine, the most commonly used biological matrices, but also hair, nails, breast milk, or adipose tissue, among others. The biomarker level has to be correlated with the health effect. One drawback of biological monitoring is that often it means an invasive method, especially when the collection of blood is required. Blood is the medium which transports chemicals and their metabolites in the body. A chemical in the blood is in dynamic equilibrium with various body tissues. Therefore, most biomarkers present in the body can be found in the blood during some period of time after exposure. Urine is still more readily and available at large volumes, allowing the determination of very low concentrations of chemicals derived from environmental exposure. However, the selection of the most suitable monitor depends on the biokinetics of the pollutant and the kind of exposure (short- or long-term) that it is intended to determine. When the site of critical action for a toxicant is known, the concentration of the biomarker at that site can be a measure of the biologically effective dose.

As aforementioned, blood and urine are the most widely used biological matrices. The advantages of blood are based on the fact that the composition is relatively constant between individuals, avoiding the need to correct the biomarker levels for individual differences. In addition, obtaining samples is straightforward (although it is an invasive procedure). Urine is more suitable for monitoring hydrophilic chemicals, metals, and metabolites than for those chemicals poorly soluble in water. The concentration of the biomarker in urine is usually correlated to its mean plasma level during the period the urine dwells in the bladder. The accuracy of the exposure estimate, using urine monitoring, depends upon the sampling strategy. The most influential factors are the time of collection and urine output. For example, measurements from 24-h specimens are more representative than from spot samples, usually better correlating with the intensity of exposure. However, collection, stabilization, and transportation of 24-h specimens in the field are difficult and often not feasible.

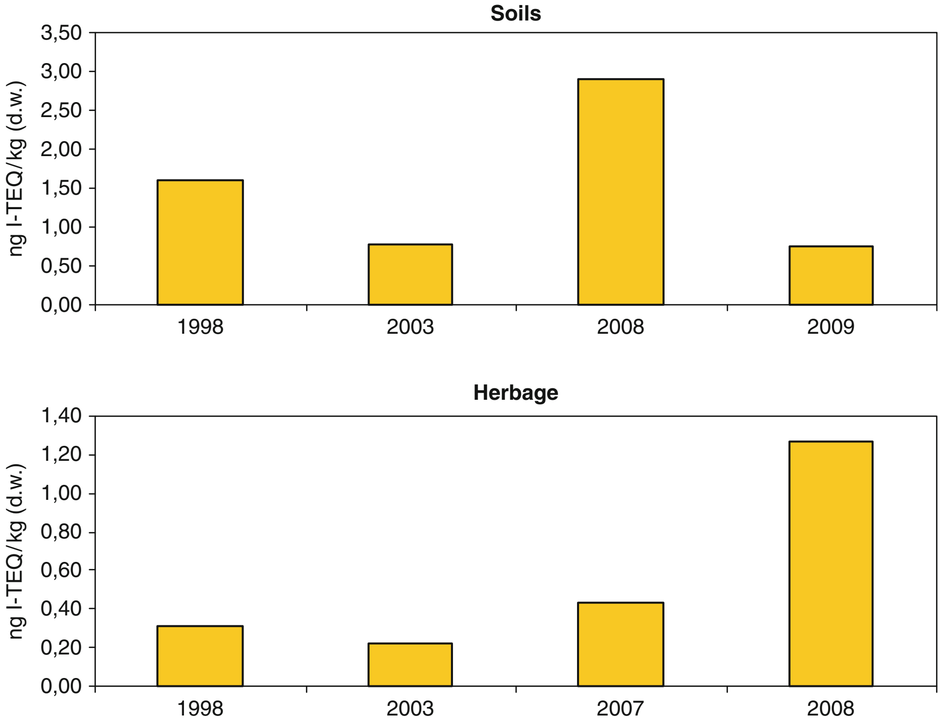
Other frequently used biological monitors are human milk and adipose tissues, which are suitable for assessing exposure to POPs such as PCDD/Fs, PCBs, and polybrominated diphenyl ethers (PBDEs). These contaminants enter humans through environmental and dietary exposure. They reach the adipose tissue, serum, and milk and equilibrate at similar levels on a fat weight basis [103]. With long (5–10 years) half-lives, POPs persist in humans as they do in the environment. On the other hand, hair is used for metal biomonitoring. In this case, exogenous contamination, the kind of hair care, etc., are factors influencing the analytical results.

A Case Study: The HWI of Constantí (Catalonia, Spain)

In this section, the study of the impact of a HWI located in Constantí (Tarragona County, Catalonia, Spain) is widely described. To determine the potential environmental impact and the human health risks derived from emissions of that HWI, environmental and biological monitoring programs were developed for metals and PCDD/Fs before the facility initiated its regular operations. These programs remain still active in order to ensure the proper performance of the facility. Some of the main results, which have been already published in international journals, are here summarized.

In 1999, the new HWI initiated its regular operations. To date, this facility is still the only one of these characteristics in Spain. It is settled in an industrial area (including a big oil refinery, a chlor-alkali plant and other chemical facilities, and a municipal solid waste incinerator). The HWI has a rotator burning furnace that operates at a temperature of 1,100°C and can burn more than 30,000 t of HW per year. The emission control system consists of an electrostatic precipitator, a catalytic converter (selective catalysis reduction) and a humid scrubber. The HWI periodically controls the PCDD/F stack emission concentrations to ensure that the cleaning systems operate properly. Moreover, to determine the potential environmental impact and the human health risks derived from emissions of the HWI, in 1996 (during the construction of the facility), a wide environmental and biological monitoring program ("background") was started and continued since then. Measurements of PCDD/Fs and heavy metals in environmental matrices collected in the vicinity of the facility have been and are being performed. Soil and herbage samples are used to assess environmental pollution, as well as to indirectly estimate the environmental exposure for the population living in the area under potential influence of the emissions of the HWI [104–107]. On the other hand, to assess human health risks, samples of various tissues, hair, plasma, and breast milk had been collected in individuals living in the vicinity of the facility [52, 108–113]. In addition, organochlorinated compounds and metals have been also monitored in blood and urine of the hypothetically most potentially exposed group of population, the HWI workers [114, 115]. All those monitoring data have been used to establish the temporal trends of PCDD/F and metal concentrations.

For the environmental program, the sampling sites were placed at different distances and wind directions from the HWI, comprehending both urban and rural areas. The concentrations of PCDD/Fs (I-TEQ) in soil and vegetation samples collected from 2003 to 2009 around the HWI of Constantí, as well as those obtained for the baseline (1998 survey; Fig. 6) showed notable fluctuations throughout the time, especially, in some specific sites. However, a general increasing tendency in PCDD/F concentrations in soils and herbage samples was not observed during the study [116]. The variations might have been caused by a number of reasons, such as the meteorological conditions occurring before the annual sampling, or the influence of local point/diffuse emission PCDD/F sources of unknown origin. This is a clear indicator that investigations of temporal trends require data collection from as many different sources as possible. Otherwise, the assessment of limited information (e.g., the use of a single compartment rather than multiple) may lead to significant misinterpretation of the current status. The environmental PCDD/F congener profiles, in both soil and vegetation, were very similar in the course of the campaigns. OCDD was the most important congener, followed by 1,2,3,4,6,7,8-HpCDD, OCDF, and 1,2,3,4,6,7,8-HpCDDF. It indicates similar significant environmental sources, even when the HWI was not operating (background levels in 1998). That entails a low impact of the HWI emissions in the surrounding environment. A particular study of the PCDD/F concentrations in environmental compartments, according to the distance to the HWI (Table 3), indicated that the highest levels of PCDD/Fs were found at the farthest distances



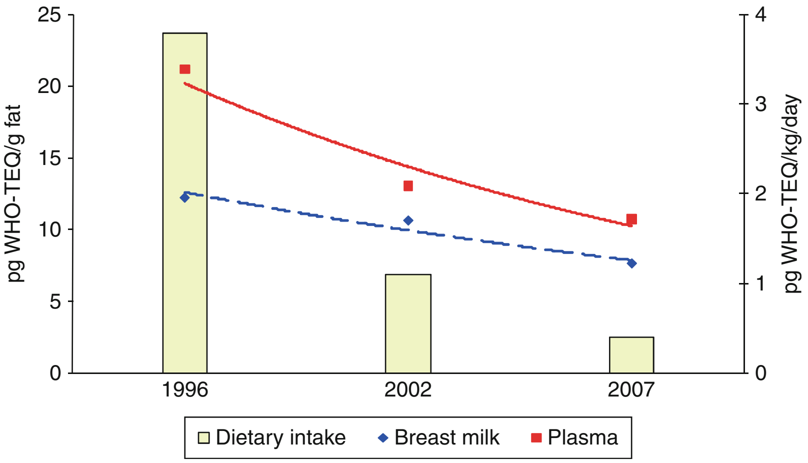
Hazardous Waste Incinerator Emissions. Figure 6
Total PCDD/F concentrations in soils and herbage collected around the HWI of Constantí (Catalonia, Spain) since 1998 (baseline study)

from the facility (>4,000 m), being 1.9- (2009) to 4.0-fold (1998) greater than those collected nearer the HWI (500 m).

With respect to the evaluation of population exposure, concentrations of PCDD/Fs and other organochlorinated compounds were determined in milk of women living in the area under direct influence of the HWI. Since it is well known that the diet is the main pathway of POP exposure for the general population, the influence of the dietary intake of these compounds on their presence in breast milk was also assessed. Milk samples were collected from women living in Tarragona County (Spain) in zones under potential influence of the HWI. The participants in the study were 25–35 years of age, who had lived in Tarragona County for at least the last 5 years. Only healthy primiparae mothers were included in the surveys. All participants completed questionnaires about age, current and former area of residence, as well as dietary habits and lifestyle. Samples were classified according to women living in urban or industrial zones. Detailed information about dietary intake of PCDD/Fs for the 1998, 2002, and 2007 surveys and their concentrations in foodstuffs were obtained from studies of our research group [117, 118]. Food samples included meats of beef (steak, hamburger), pork (loin, sausage), chicken (breast), lamb (steak), and meat products (ham, hot dogs, salami); fish (hake, sardine), shellfish (mussel), and tinned fish (tuna, sardine); vegetables (lettuce, tomato, potato, green beans, cauliflower); fresh fruits (apple, orange, pear); eggs; cow milk (whole, semi-skimmed) and dairy products (yogurt, cheese); cereals (bread, pasta, rice); pulses (lentils, beans); fats (margarine) and oils (olive, sunflower). Composites of 10 individual samples were pooled and analyzed for PCDD/Fs. Average daily food consumption data were obtained from studies carried out in the same area [117]. PCDD/F mean concentrations in the breast milk of women living in the area of

Hazardous Waste Incinerator Emissions. Table 3 Summary of PCDD/F concentrations (ng I-TEQ/kg dry weight) in soil and herbage samples according to the distance from the HWI of Constantí (Spain)

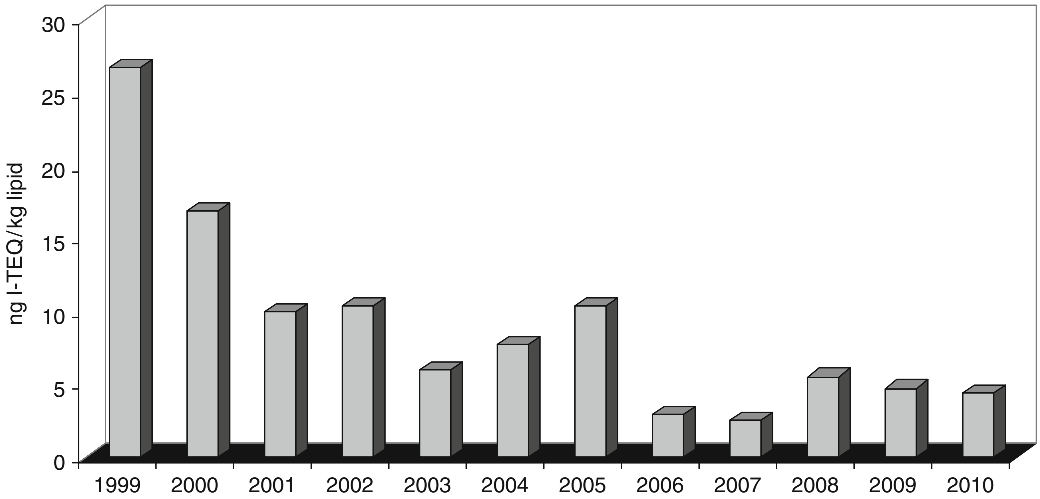
Monitor	Year	500 m	3,000 m	>4,000 m
Soil	1998	0.83 ± 0.41	0.75 ± 0.29	3.36 ± 5.03
	2003	0.59 ± 0.36	0.40 ± 0.30	1.28 ± 1.20
	2008	0.39 ± 0.09	13.2 ± 25.0	1.41 ± 1.73
	2009	0.66 ± 0.78	1.13 ± 1.27	1.23 ± 1.16
Herbage	1998	0.29 ± 0.16	0.21 ± 0.04	0.49 ± 0.58
	2003	0.20 ± 0.06	0.18 ± 0.08	0.22 ± 0.10
	2007	0.27 ± 0.09	0.32 ± 0.15	0.78 ± 0.56
	2008	0.57 ± 0.55	1.45 ± 1.27	1.77 ± 0.77



Hazardous Waste Incinerator Emissions. Figure 7
Temporal trends of PCDD/Fs in human plasma and breast milk, as well as comparison with dietary exposure to PCDD/Fs in studies performed in 1998 (baseline), 2002 and 2007

study were 12.2, 10.6, and 7.6 pg WHO-TEQ/g lipid in 1998, 2002, and 2007, respectively (Fig. 7). The historical trend of PCDD/Fs in breast milk of women living in the neighborhood of the HWI was in agreement with the continued decline observed in the emissions and atmospheric concentrations of POPs [119]. The dietary intakes of PCDD/Fs for the adult population were 210 pg I-TEQ/day, 63.80 pg WHO-TEQ/day, and 27.81 pg WHO-TEQ/day in 1998, 2002, and 2007, respectively. Furthermore, the decrease in the dietary intake of these pollutants was higher than the reduction observed in biological tissues (blood and milk), making evident that the effect of the decrease in the dietary intake of PCDD/Fs is not immediately reflected in a reduction of the PCDD/F concentrations in human plasma or milk [103]. Food contamination is primarily caused by airborne POPs that settle on plants and water and accumulate in the food chain, where they are particularly associated with fat. In the program here reported, and taking into account that the HWI is placed in an industrial zone without farms and cattle in the surroundings, it is important to remark that almost all foods consumed in the area come from a diversity of origins, without any direct relationship with the pollutants emitted by the HWI. It is also important to highlight that in recent years, the dietary habits in young women in Spain have changed.

Concerning metals, especially in environmental monitors, the results of a great number of studies, and also in this specific program performed during the last 10 years, show that it is quite usual to find fluctuations in the concentrations of metals through the monitoring period. As a result, if possible, it is important to combine the use of different matrices and maintain the study through a period of time long enough to detect possible changes and trends. It must be taken into account that in addition to combustion processes, environmental levels of metals can be notably influenced by a remarkable number of different emission sources. The HWI is located in an industrial zone, where a considerable number of chemical and petrochemical companies are settled in. Moreover, the area is surrounded by a highway and roads with a heavy traffic. Consequently, the potential impact of other sources must be specially considered. On the other hand, no significant correlation (Pearson) between soil and herbage concentrations was found for any of the analyzed elements. It means that root uptake of heavy metals by plants can be considered negligible. Likewise, the lack of a homogeneous change between soil and vegetation levels also means that other potential emission sources, different from incineration, could be even more important than in principle expected. To assess the influence of the HWI on the



Hazardous Waste Incinerator Emissions. Figure 8
Temporal trends of concentrations of PCDD/Fs in plasma of workers at the HWI of Constantí (Spain) since the plant started its regular operations

area under its direct influence, sampling points collected at different distances and wind directions were grouped and analyzed for clustering. Only significant differences in metal concentrations appeared at distances >1.5 km from the plant, being an indicator that the HWI is not the main contributor of the presence of metals into the environmental area under evaluation. Data fluctuations would be due to the presence of other pollution sources (traffic, industrial activities, etc.) and/or to the environmental variations of these elements in the zone.

With regard to human exposure, the internal exposure of workers of the HWI to a number of organic substances such as PCDD/Fs has been assessed from the beginning of the activity. Information concerning health risks of workers in waste incinerators is rather scarce. In addition, most of these studies are more focused on workers in MSW incinerators (MSWI) than in those working at HWI. Baseline levels of these substances were determined in the future workers of the evaluated HWI during the period of construction of the facility, and since then, they have been annually performed. Figure 8 depicts the temporal trends of PCDD/F concentrations in plasma of the workers of the HWI [120]. A remarkable decrease was noted with respect to the baseline levels. As already commented, it is well established that for nonoccupationally exposed population, the main pathway of exposure to PCDD/Fs is the diet. Therefore, the results observed in the plasma levels of PCDD/Fs of the workers of the HWI indicated that not only the occupational exposure was not significant, but also an important reduction in the dietary intake during the last years, as it has been previously stated.

Future Directions

Hazardous waste incinerators might be a significant source of a wide range of pollutants if stack gases are not properly treated. In fact, in the past, they were a leading source of dioxin emissions to the environment. The composition of the emissions depends on the composition of the materials burned and also on the combustion conditions and air pollution devices used to clean them. Although PCDD/Fs and metals are found at rather low concentrations in comparison with other pollutants, they are among the contaminants that raise more concern for the population and public opinion because of their persistence, toxicity, and capacity of bioaccumulation. When the burning conditions and the APCD used are suitable, pollutant emissions from HWIs are low. However, when incinerators are operating in ways not intended in their original design, emissions may be then significant. Moreover, the emissions of some pollutants, especially PCDD/Fs, vary when operating conditions are not stable (e.g., star up). Legislative regulations indicate that emissions of PCDD/Fs from waste incinerators must be monitored once or twice a year, and when operating conditions are stable. However, that means an underestimation of the real emissions because in unstable conditions emissions are higher. Recent advances in the methodologies for stack emissions measurements of PCDD/Fs already allow measuring their concentrations in continuous. However, few studies have investigated source samples. Instead of that, environmental and biological monitoring is frequently carried out to evaluate the impact of pollutant sources and human health risks. Environmental monitoring programs are common tools to control the state of pollution in a particular area. These programs are used to assess the contribution of a potential contaminant source on the global environment. Chemical pollutants are determined in environmental matrices as air, water, soil, and food. Human biological monitoring has become an important tool in environmental and occupational medicine for assessment of the levels of internal exposure to harmful substances taken up from the environment. The selection of the most suitable biomonitors depends on the biokinetics of the pollutant and the kind of exposure (short- or long- time), which should be determined. Although blood and urine are the most common, others like breast milk, hair, etc., are also commonly used.

The environmental and human impact of the HWI located in Constantí (Spain) has been evaluated since it started to regularly operate in 1999. The concentrations of heavy metals and PCDD/Fs have been determined in environmental monitors around the facility and also in biomonitors of the nearby population, as well as in workers. So far, the results indicate that the emissions of the HWI do not mean a significant source of PCDD/Fs and metal pollution in the area. However, some results, especially those of heavy metals in environmental monitors show sometimes certain fluctuations. It is a clear indicator that investigations of temporal trends require data collection from as many different sources as possible. Otherwise, the assessment of limited information might lead to significant misinterpretation of the current situation. The global PCDD/F pattern analysis of the environmental and biological samples have shown that the weight of the PCDD/F stack emissions of the HWI on the environmental burden and on exposure of the individuals living in the surroundings of the facility is not significant in relation to the background levels. These results are in agreement with those of other of recent studies where the repercussion of incinerators, especially those in the EU, has been evaluated. These investigations clearly state that waste-to-energy plants using BAT and BEP to meet the stringent emission standards, result in a high reduction in expected risks with respect to older plants with no specific PCDD/F control measures [120]. Therefore, it is clear that if waste incineration plants are equipped with BAT, the presence of these facilities should not mean additional human health risks of exposure to PCDD/Fs (and metals). In the EU, the current Directive 2008/98/EC or Waste Framework Directive (WFD) sets the basic concepts and definitions related to waste management, proposing a hierarchy in the alternatives for waste treatment. In this prioritization, incineration involving the recovery of energy (or waste-to-energy) is enhanced, after prevention, reuse, and recycling of waste. In this context, waste management planning includes the future construction of new incineration plants, some of them for the management of HW. Although this kind of facilities are constrained to follow strict regulations regarding the emissions of chemical pollutants, such as dioxins and mercury, a special control of the air quality around incineration plants must follow through the execution of environmental monitoring programs to assure that their operative activities do not pose a risk for the public health and the environment in a long term.

Bibliography

1. UNEP (2009) Developing integrated solid waste management plan, vol 2: assessment of current waste management system and gaps therein. United Nations Environmental Programme Division of Technology, Industry and Economics International Environmental Technology Centre, Osaka/Shiga
2. Rivera-Austrui J, Borrajo M, Martinez K, Adrados M, Abalos M, Van Bavel B, Rivera J, Abad E (2011) Assessment of polychlorinated dibenzo-p-dioxin and dibenzofuran emissions from a hazardous waste incineration plant using long-term sampling equipment. Chemosphere 82:1343–1349
3. Olie K, Vermeulen PL, Hutzinger O (1977) Chlorodibenzo-*p*-dioxins and chlorodibenzofurans are trace components of fly ash and flue gas of some municipal incinerators in the Netherlands. Chemosphere 8:455–459
4. UNILABS Environmental (2001) Characterisation and estimation of dioxin and furan emissions from waste incineration facilities. www.environment.gov.au/settlements/publications/chemicals/dioxins/dioxinemissions.html
5. Gullett BK, Touati A, Lee C (2000) Formation of chlorinated dioxins and furans in a hazardous-waste-firing industrial boiler. Environ Sci Technol 34:2069–2074
6. Choi K-I, Lee S-H, Lee D-H (2008) Emissions of PCDDs/DFs and dioxin-like PCBs from small waste incinerators in Korea. Atmos Environ 42:940–948
7. Rimmer DL, Vizard CG, Pless-Mulloli T, Singleton I, Air VS, Keatinge ZA (2006) Metal contamination of urban soils in the vicinity of a municipal waste incinerator: one source among many. Sci Total Environ 356:207–216
8. Jung CH, Matsuto T, Tanaka N, Okada, T (2004) Metal distribution in incineration residues of municipal solid waste (MSW) in Japan. Waste Manage 24:381–391
9. Donnelly JR (1992) Metal emissions control technologies for waste incineration. In: Series AS (ed) Clean energy from waste and coal. American Chemical Society, Washington, DC, pp 174–188
10. Kim JH, Seo YC, Pudasainee D, Lee SH, Cho SJ, Jang HN, Park JM, Song GJ, Park KS (2009) Efforts to develop regulations in Korea similar to the US maximum achievable control technology (MACT) regulations for hazardous waste incinerators. J Mat Cycl Waste Manage 11(3):183–190
11. Kutz FW, Barnes DG, Bottimore DP, Greim H, Bretthauer EW (1990) The international toxicity equivalency factor (I-TEF) method of risk assessment for complex mixtures of dioxins and related compounds. Chemosphere 20:751–757
12. Van den Berg M, Birnbaum LS, Denison M, De Vito M, Farland W, Feeley M, Fiedler H, Hakansson H, Hanberg A, Haws L, Rose M, Safe S, Schrenk D, Tohyama C, Tritscher A, Tuomisto J, Tysklind M, Walker N, Peterson RE (2006) The 2005 World Health Organization reevaluation of human and Mammalian toxic equivalency factors for dioxins and dioxin-like compounds. Toxicol Sci 93(2):223–241
13. NATO (1988) International toxicity equivalence factors (I-TEF) method of risk assessment for complex mixtures of dioxins and related compounds. Pilot study on international information exchange on dioxins and related compounds, Report number 176, Aug 1988. North Atlantic Treaty Organisation, Committee on Challenges of Modern Society
14. Kimbrough RD (1995) Polychlorinated biphenyls (PCBs) and human health: an update. Crit Rev Toxicol 25:133–163

15. ATSDR (2000) Polychlorinated biphenyls. Agency for Toxic Substances and Disease Registry, Atlanta
16. Aoki Y (2001) Polychlorinated biphenyls, polychlorinated dibenzo-*p*-dioxins, and polychlorinated dibenzofurans as endocrine disrupters – what we have learned from Yusho disease. Environ Res 86:2–11
17. IARC (1995) Monograph volumes 1972–1994. WHO, Geneva, pp 1–60. www.holgr/medical/nuclear/carcinogen.htm
18. Negri E, Bosetti C, Fattore E, La Vecchia C (2003) Environmental exposure to polychlorinated biphenyls (PCBs) and breast cancer: a systematic review of the epidemiological evidence. Eur J Cancer Prev 12:509–516
19. Helm PA, Bidleman TF (2003) Current combustion-related sources contribute to polychlorinated naphthalene and dioxin-like polychlorinated biphenyl levels and profiles in air in Toronto, Canada. Environ Sci Technol 37:1075–1082
20. Meijer SN, Harner T, Helm PA, Halsall CJ, Johnston AE, Jones KC (2001) Polychlorinated naphthalenes in U.K. soils: time trends, markers of source, and equilibrium status. Environ Sci Technol 35:4205–4213
21. Lin Y, Gupta G, Baker J (1995) Photodegradation of polychlorinated biphenyl congeners using simulated sunlight and diethylamine. Chemosphere 31:3323–3344
22. Gouin T, Cousins I, Mackay D (2004) Comparison of two methods for obtaining degradation half-lives. Chemosphere 56:531–535
23. Sinkkonen S, Paasivirta J (2000) Degradation half-life times of PCDDs, PCDFs and PCBs for environmental fate modeling. Chemosphere 40:943–949
24. Shirai JH, Kissel JC (1996) Uncertainty in estimated half-lives of PCBs in humans: impact on exposure assessment. Sci Total Environ 187:199–210
25. Brown JF (1994) Determination of PCB metabolic, excretion, and accumulation rates for use as indicators of biological response and relative risk. Environ Sci Technol 28:2295–2305
26. Nizzetto L, Cassani C, Di Guardo A (2006) Deposition of PCBs in mountains: the forest filter effect of different forest ecosystem types. Ecotoxicol Environ Safe 63:75–83
27. Falandysz J (1998) Polychlorinated naphthalenes: an environmental update. Environ Pollut 101:77–90
28. Blankenship AL, Kannan K, Villalobos SA, Villeneuve DL, Falandysz J, Imagawa T, Jakobsson E, Giesy JP (2000) Relative potencies of individual polychlorinated naphthalenes and halowax mixtures to induce Ah receptor-mediated responses. Environ Sci Technol 34:3153–3158
29. Hayward D (1998) Identification of bioaccumulating polychlorinated naphthalenes and their toxicological significance. Environ Res 76:1–18
30. Van den Berg M, Peterson RE, Schrenk D (2000) Human risk assessment and TEFs. Food Addit Contam 17:347–358
31. Kannan K, Hilscherova K, Imagawa T, Yamashita N, Williams LL, Giesy JP (2001) Polychlorinated naphthalenes, -biphenyls, -dibenzo-*p*-dioxins, and -dibenzofurans in double-crested cormorants and herring gulls from Michigan waters of the Great Lakes. Environ Sci Technol 35:441–447
32. Eljarrat E, Barcelo D (2003) Priority lists for persistent organic pollutants and emergingcontaminants based on their relative toxic potency in environmental samples. Trend Anal Chem 22:655–665
33. Jarnberg U, Asplund L, De Wit C, Grafstrom AK, Haglund P, Jansson B, Lexen K, Strandell M, Olsson M, Jonsson B (1993) Polychlorinated biphenyls and polychlorinated naphthalenes in Swedish sediment and biota: levels, patterns, and time trends. Environ Sci Technol 27:1364–1374
34. Schneider M, Stieglitz L, Will R, Zwick G (1998) Formation of polychlorinated naphthalenes on fly ash. Chemosphere 37:2055–2070
35. Abad E, Caixach J, Rivera J (1999) Dioxin like compounds from municipal waste incinerator emissions: assessment of the presence of polychlorinated naphthalenes. Chemosphere 38:109–120
36. Kannan K, Imagawa T, Blankenship AL, Giesy JP (1998) Isomer-specific analysis and toxic evaluation of polychlorinated naphthalenes in soil, sediment, and biota collected near the site of a former chlor-alkali plant. Environ Sci Technol 32: 2507–2514
37. Yamashita N, Kannan K, Imagawa T, Miyazaki A, Giesy JP (2000) Concentrations and profiles of polychlorinated naphthalene congeners in eighteen technical polychlorinated biphenyl preparations. Environ Sci Technol 34:4242–4254
38. Meijer SN, Ockenden WA, Steinnes E, Corrigan BP, Jones KC (2003) Spatial and temporal trends of POPs in Norwegian and UK background air: Implications for global cycling. Environ Sci Technol 37:454–461
39. Harner T, Bidleman TF (1998) Measurement of octanol-air partition coefficients for polycyclic aromatic hydrocarbons and polychlorinated naphthalenes. J Chem Eng Data 43: 40–46
40. Domingo JL (2004) Polychlorinated naphthalenes in animal aquatic species and human exposure through the diet: a review. J Chromat A 1054:327–334
41. Opperhuizen A, van der Velde EW, Gobas FAPC, Liem DAK, van der Steen JMD (1985) Relationship between bioconcentration in fish and steric factors of hydrophobic chemicals. Chemosphere 14:1871–1896
42. Hanari N, Horii Y, Taniyasu S, Falandysz J, Bochentin I, Orlikowska A, Puzyn T, Yamashita N (2004) Isomer specific analysis of polychlorinated naphthalenes in pine trees (*Pinus thunbergi* Parl.) and (*Pinus densiflora* Sieb. et Zucc) needles around Tokyo Bay, Japan. Pol J Environ Stud 13:139–151
43. Crookes MJ, Howe PP (1993) Environmental hazard assessment: halogenated naphthalenes. Department of the Environment, Directorate for Air, Climate and Toxic Substances, Toxic Substances Division, Garston
44. Van de Plassche E, Schwegler A, Iestra W (2002) Polychlorinated naphthalenes and the UN-ECE POP protocol. Organohal Compd 58:89–91
45. Breivik K, Alcock R, Li Y-F, Bailey RE, Fierens S, Pacyna JM (2004) Primary sources of selected POPs: regional and global scale emission inventories. Environ Pollut 128:3–16

46. Lönnermark A, Blomqvist P, Marklund S (2008) Emissions from simulated deep-seated fires in domestic waste. Chemosphere 70:626–639
47. ATSDR (2002) Toxicological profile for hexachlorobenzene. U.S. Public Health Services, U.S. Department of Health and Human Services, Atlanta
48. Messerer P, Zober A, Becher H (1998) Blood lipid concentrations of dioxins and furans in a sample of BASF employees included in the IARC registry of workers exposed to phenoxy acid herbicides and/or chlorophenols. Environ Health Perspect 106:733–735
49. ATSDR (2000) Toxicological profile for arsenic. US Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry, Atlanta
50. Basu A, Mahata J, Gupta S, Giri AK (2001) Genetic toxicology of a paradoxical human carcinogen, arsenic: a review. Mutat Res 488:171–194
51. Pongratz R (1998) Arsenic speciation in environmental samples of contaminated soil. Sci Total Environ 224:133–141
52. Nadal M, Perelló G, Schuhmacher M, Cid J, Domingo JL (2008) Concentrations of PCDD/PCDFs in plasma of subjects living in the vicinity of a hazardous waste incinerator: follow-up and modeling validation. Chemosphere 73:901–906
53. Fattorini D, Alonso-Hernandez CM, Diaz-Asencio M, Munoz-Caravaca A, Pannacciulli FG, Tangherlini M, Regoli F (2004) Chemical speciation of arsenic in different marine organisms: Importance in monitoring studies. Marine Environ Res 58:845–850
54. IARC (1987) Overall evaluations of carcinogenicity: an updating of IARC monographs. In: IARC Monographs on the evaluation of carcinogenic risks to humans, vol 1–42(suppl 7). International Agency for Research on Cancer, Lyon
55. Rossman TG, Uddin AN, Burns FJ (2004) Evidence that arsenite acts as a cocarcinogen in skin cancer. Toxicol Appl Pharmacol 198:394–404
56. Huff J, Chan P, Nyska A (2000) Is the human carcinogen arsenic carcinogenic to laboratory animals? Toxicol Sci 55:17–23
57. ATSDR (2002) Toxicological profile for beryllium. Department of Health and Human Services, Atlanta
58. US EPA (2010) Preliminary remediation goals. www.epa.gov/region9/superfund/prg/. Accessed 13 Apr 2011
59. Baldwin DR, Marshall WJ (1999) Heavy metal poisoning and its laboratory investigation. Ann Clin Biochem 36:267–300
60. Robards K, Worsford P (1991) Cadmium: toxicology and analysis. A review. Analyst 116:549–568
61. McElroy JA, Shafer MM, Trentham-Dietz A, Hampton JM, Newcomb PA (2006) Cadmium exposure and breast cancer risk. J Nat Cancer Inst 98:869–873
62. Kasuya MR (2000) Recent epidemiological studies on itai-itai disease as a chronic cadmium poisoning in Japan. Water Sci Technol 42:147–154
63. Satarug S, Moore MR (2004) Adverse health effects of chronic exposure to Low-level cadmium in foodstuffs and cigarette smoke. Environ Health Perspect 112:1099–1103
64. Linna A, Oksa P, Palmroos P, Roto P, Laippala P, Uitti J (2003) Respiratory health of cobalt production workers. Am J Ind Med 44:124–132
65. ATSDR (2004) Toxicological profile for cobalt. Department of Health and Human Services, Atlanta
66. ATSDR (2004) Toxicological profile for copper. Department of Health and Human Services, Atlanta
67. Singh J, Carlisle DL, Pritchard DE, Patierno SR (1998) Chromium-induced genotoxicity and apoptosis: relationship to chromium carcinogenesis (review). Oncol Rep 5:1307–1318
68. Vaglenov A, Nosko M, Georgieva R, Carbonell E, Creus A, Marcos R (1999) Genotoxicity and radioresistance in electroplating workers exposed to chromium. Mutat Res 446(1):23–34
69. ATSDR (2000) Toxicological profile for chromium. US Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry, Atlanta
70. IARC (1990) Chromium and chromium compounds. In: IARC Monographs on the evaluation of carcinogenic risks to humans, vol 49. International Agency for the Research on Cancer, Lyon
71. ATSDR (2005) Toxicological profile for tin. Department of Health and Human Services, Atlanta
72. ATSDR (2000) Toxicological profile for manganese. US Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry, Atlanta
73. Josephs KA, Ahlskog JE, Klos KJ, Kumar N, Fealey RD, Trenerry MR, Cowl CT (2005) Neurologic manifestations in welders with pallidal MRI T1 hyperintensity. Neurology 64:2033–2039
74. ATSDR (1999) Toxicological profile for mercury. US Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry, Atlanta
75. Chang MB, Wu HT, Huang CK (2000) Evaluation on speciation and removal efficiencies of mercury from municipal solid waste incinerators in Taiwan. Sci Total Environ 246:165–173
76. Karimi A, Moniri F, Nasihatkon A, Zarepoor MJ, Alborzi A (2002) Mercury exposure among residents of a building block in Shirz, Iran. Environ Res A 88:41–43
77. Ekino S, Susa M, Ninomiya T, Imamura K, Kitamura T (2007) Minamata disease revisited: an update on the acute and chronic manifestations of methyl mercury poisoning. J Neurol Sci 262:131–144
78. Horng CJ, Tsai JL, Lin SR (1999) Determination of urinary arsenic, mercury, and selenium in steel production workers. Biol Trace Elem Res 70:29–40
79. ATSDR (2005) Toxicological profile for nickel. U.S. Public Health Services, U.S. Department of Health and Human Services, Atlanta
80. Finkelstein Y, Markowitz ME, Rosen JF (1998) Low-level lead-induced neurotoxicity in children: an update on central nervous system effects. Brain Res Rev 27:168–176
81. Torrente M, Colomina MT, Domingo JL (2005) Metal concentrations in hair and cognitive assessment in an adolescent population. Biol Trace Elem Res 104:215–221

82. Navas-Acien A, Schwartz BS, Rothenberg SJ, Hu H, Silbergeld EK, Guallar E (2008) Bone lead levels and blood pressure endpoints: a meta-analysis. Epidemiology 19:496–504
83. dos Santos ACD, Colacciopo S, Dal Bó CMR, dos Santos NAG (1994) Occupational exposure to lead, kidney function tests, and blood pressure. Am J Ind Med 26:635–645
84. Saha A (2005) Thallium toxicity: a growing concern. Indian J Occup Environ Med 9:53–56
85. ATSDR (1992) Toxicological profile for thallium. Department of Health and Human Services, Atlanta
86. ATSDR (1992) Toxicological profile for vanadium. Department of Health and Human Services, Atlanta
87. Hinshaw GD, Trenholm AR (2001) Hazardous waste incineration in perspective. Waste Manage 21:471–475
88. van Velzen D, Langenkamp H, Herb G (2002) Review: mercury in waste incineration. Waste Manage Res 20:556–568
89. Löthgren C-J, van Bavel B (2005) Dioxin emissions after installation of a polishing wet scrubber in a hazardous waste incineration facility. Chemosphere 61(3):405–412
90. Van Caneghem J, Block C, Van Brecht A, Wauters G, Vandecasteele C (2010) Mass balance for POPs in hazardous and municipal solid waste incinerators. Chemosphere 78: 701–708
91. Kim B-H, Lee S-J, Mun S-J, Chang Y-S (2005) A case study of dioxin monitoring in and around an industrial waste incinerator in Korea. Chemosphere 58:1589–1599
92. Morselli L, Passarini F, Bartoli M (2002) The environmental fate of heavy metals arising from a MSW incineration plant. Waste Manage Res 22:875–881
93. Ahammed AKMR, Nixon BM (2006) Environmental impact monitoring in the EIA process of South Australia. Environ Impact Assess Rev 26(5):426–447
94. Klumpp A, Klumpp G, Domingos M (1994) Plants as bioindicators of air pollution at the Serra do Mar near the industrial complex of Cubatao, Brazil. Environ Pollut 85:109–116
95. Moraes RM, Klumpp A, Furlan CM, Klumpp G, Domingos M, Rinaldi MC, Modesto IF (2002) Tropical fruit trees as bioindicators of industrial pollution in southeast Brazil. Environ Int 28:367–374
96. Ho YB, Tai KM (1988) Elevated levels of lead and other metals in roadside soil and grass and their use to monitor aerial metal depositions in Hong Kong. Environ Pollut 49:37–51
97. Meneses M, Llobet JM, Granero S, Schuhmacher M, Domingo JL (1999) Monitoring metals in the vicinity of a municipal waste incinerator: temporal variation in soils and vegetation. Sci Total Environ 226(2–3):157–164
98. Nadal M, Schuhmacher M, Domingo JL (2004) Metal pollution of soils and vegetation in an area with petrochemical industry. Sci Total Environ 321:59–69
99. Ewers U, Krause C, Schulz C, Wilhelm M (1999) Reference values and human biological monitoring values for environmental toxins. Report on the work and recommendations of the commission on human biological monitoring of the German Federal Environmental Agency. Int Arch Occup Environ Health 72:255–260
100. Paustenbach D, Galbraith D (2006) Biomonitoring: is body burden relevant to public health? Regul Toxicol Pharmacol 44:249–261
101. Calafat AM, Ye X, Silva MJ, Kuklenyik Z, Needham LL (2006) Human exposure assessment to environmental chemicals using biomonitoring. Int J Androl 29:166–171
102. Angerer J, Ewers U, Wilhelm M (2007) Human biomonitoring: state of the art. Int J Hyg Environ Health 210:201–228
103. Schuhmacher M, Kiviranta H, Ruokojärvi P, Nadal M, Domingo JL (2009) Concentrations of PCDD/Fs, PCBs and PBDEs in breast milk of women from Catalonia, Spain: a follow-up study. Environ Int 35:607–613
104. Schuhmacher M, Agramunt MC, Rodriguez-Larena MC, Díaz-Ferrero J, Domingo JL (2002) Baseline levels of PCDD/Fs in soil and herbage samples collected in the vicinity of a new hazardous waste incinerator in Catalonia, Spain. Chemosphere 46:1343–1350
105. Schuhmacher M, Rodriguez-Larena MC, Agramunt MC, Diaz-Ferrero J, Domingo JL (2002) Environmental impact of a new hazardous waste incinerator in Catalonia, Spain: PCDD/PCDF levels in herbage samples. Chemosphere 48:187–193
106. Mari M, Nadal M, Ferré-Huguet N, Schuhmacher M, Borrajo M, Domingo JL (2007) Monitoring PCDD/Fs in soil and herbage samples collected near a hazardous waste incinerator. Health risks for the population living nearby. Human Ecol Risk Assess 13:1255–1270
107. Nadal M, Bocio A, Schuhmacher M, Domingo JL (2005) Trends in the levels of metals in soils and vegetation samples collected near a hazardous waste incinerator. Arch Environ Contam Toxicol 49:290–298
108. Nadal M, Domingo JL, García F, Schuhmacher M (2009) Levels of PCDD/F in adipose tissue on non-occupationally exposed subjects living near a hazardous waste incinerator in Catalonia, Spain. Chemosphere 74:1471–1476
109. Nadal M, Espinosa G, Schuhmacher M, Domingo JL (2004) Patterns of PCDDs and PCDFs in human milk and food and their characterization by artificial neural networks. Chemosphere 54:1375–1382
110. Schuhmacher M, Domingo JL, Hagberg J, Lindström G (2004) PCDD/F and non-ortho PCB concentrations in adipose tissue of individuals living in the vicinity of a hazardous waste incinerator. Chemosphere 57:357–364
111. Schuhmacher M, Domingo JL, Llobet JM, Kiviranta H, Vartiainen T (1999) PCDD/F concentrations in milk of nonoccupationally exposed women living in southern Catalonia, Spain. Chemosphere 38:995–1004
112. Schuhmacher M, Domingo JL, Llobet JM, Lindström G, Wingfors H (1999) Dioxin and dibenzofuran concentrations in adipose tissue of a general population from Tarragona, Spain. Chemosphere 38:2475–2487
113. Schuhmacher M, Domingo JL, Llobet JM, Lindström G, Wingfors H (1999) Dioxin and dibenzofuran concentrations in blood of a general population from Tarragona, Spain. Chemosphere 38:1123–1133

114. Mari M, Borrajo MA, Schuhmacher M, Domingo JL (2007) Monitoring PCDD/Fs and other organic substances in workers of a hazardous waste incinerator: a case study. Chemosphere 67:574–581
115. Carmen Agramunt M, Domingo A, Domingo JL, Corbella J (2003) Monitoring internal exposure to metals and organic substances in workers at a hazardous waste incinerator after 3 years of operation. Toxicol Lett 146:83–91
116. Vilavert L, Nadal M, Mari M, Schuhmacher M, Domingo JL (2010) Monitoring temporal trends in environmental levels of polychlorinated dibenzo-*p*-dioxins and dibenzofurans: results from a 10-year surveillance program of a hazardous waste incinerator. Arch Environ Contam Toxicol 59:521–531
117. Martí-Cid R, Bocio A, Domingo JL (2008) Dietary exposure to PCDD/PCDFs by individuals living near a hazardous waste incinerator in Catalonia, Spain: temporal trend. Chemosphere 70:1588–1595
118. Bocio A, Domingo JL (2005) Daily intake of polychlorinated dibenzo-p-dioxins/polychlorinated dibenzofurans (PCDD/PCDFs) in foodstuffs consumed in Tarragona, Spain: a review of recent studies (2001–2003) on human PCDD/PCDF exposure through the diet. Environ Res 97:1–9
119. Hassanin A, Lee RGM, Johnston AE, Jones KC (2006) Reductions and changing patterns of ambient PCDD/Fs in the UK: evidence and implications. Chemosphere 65: 530–539
120. Mari M, Schuhmacher M, Domingo JL (2009) Levels of metals and organic substances in workers at a hazardous waste incinerator: a follow-up study. Int Arch Occup Environ Health 82:519–528

Hazardous Waste, Introduction

MERVIN FINGAS
Spill Science, Edmonton, AB, Canada

Hazardous waste refers to any waste material that contains toxic or potentially toxic substances. These toxic substances may range from mildly toxic up to highly toxic and may even include highly radioactive materials (see ► Hazardous Materials Characterization and Assessment). The first important definition of a hazardous material mix or stream is an exacting definition of a particular waste stream, its variances and potential contents. The variances in the hazardous content also constitute an important characteristic of the waste. In addition, such waste streams often have measurements such as toxicity testing and other characterization by various testing methods. The hazards considered in hazardous waste are many, including physical hazards such as explosive characteristics, health hazards, and environmental hazards.

Industrial hazardous waste is also a topic here and is often thought of as the true hazardous waste. In reality, this waste is often mixed with domestic waste. Sometimes as well, industrial hazardous waste is treated as a separate waste stream. Industrial wastes are often subject to specific legislation as regards disposal. Often, specific waste streams are subject to detailed legislation regarding disposal. Items of this type include batteries, electronic equipment, mercury-containing products, heavy-metal-containing products, ash from incinerators or crematoriums, refrigerants, etc.

The methods of dealing with hazardous waste are as varied as the waste itself. This section has been divided into incineration methods and non-incineration methods. The incineration methods are very important in terms of low-level hazardous waste, especially that containing municipal waste. Non-incineration methods include landfilling, a method not covered in this encyclopedia, vitrification, solidification/stabilization, specific treatment methodologies, and specific product recycling (► Non-Flame Incineration).

Incineration has developed to a great extent in the past generation. In the past, the problems with incineration, particularly those of the gaseous or stack emissions, and the problems with ash disposal were very limiting to the technology for hazardous waste. Extensive development has led to the better control of these emissions and provided means to deal with incinerator ash. More about this will be discussed later. Incineration is particularly used at this time to deal with organic wastes such as contaminated municipal waste, fuels, and combustible streams.

Related methods include pyrolysis, the heating of materials to high temperatures, without combustion, and thermal desorption. The latter are heating of waste to high temperatures to remove volatiles such as gasoline and diesel fuel. This is now used frequently to deal with soils contaminated with fuels. These methods are still largely under development at this time, however, have specific applicability.

Non-incineration methods include landfilling. Landfilling is the traditional method of dealing with low-level hazardous waste but has rapidly fallen out of favor because of leaching and other problems. The modern use of engineered landfills has changed this somewhat. Engineered landfills include liners to avoid leaching, surface cover to avoid passage of water through the fill, and relief or recovery of gases produced by the decomposing fill.

Of course, among the non-incineration methods is the desirable method of recycling or partial recycling. This is less applicable to hazardous waste than to domestic waste; however, it is increasingly being invoked. Among the materials, solvents, metals, and similar materials are increasingly being recovered.

Another consideration, especially applicable to incineration, is the recovery of energy. Many incineration plants now recover the heat from combustion and convert it to electricity. This is, of course, applicable to hazardous waste that contains sufficient amount of combustible materials. The concept of energy recovery is most often applied to domestic waste which typically contains sufficient caloric value to be economically recovered.

Vitrification is another technology that is useful for dealing with certain types of hazardous materials (▶ Vitrification of Waste and Reuse of Waste-derived Glass). Amorphous glasses form from the oxides of many materials, especially silica. Vitrification is sometimes applied to metal oxides and radioactive materials. These hazardous waste materials can be added to glasses to form a stable new glass. This material is typically disposed of rather than used in any manner. Such glasses can be stable for millennia.

Incineration continues to be a major disposal mechanism for hazardous wastes (▶ Incineration Technologies). The issues with incineration continue to be a topic of constant research. These issues include emissions and formation of new compounds. The control factors for incineration include oxygenation, temperature, and residence time. These factors, in turn, strongly influence emission type and amount. Most incinerators include several stages of flue gas cleaning to minimize emissions. Many incinerators have residue or ash conditioning to ensure that these materials can safely be disposed of. There are many types of incinerators, and some are finding increasing utility in the disposal of hazardous materials. Fluidized bed incinerators are very popular and are able to handle many types of wastes. Rotary-kiln incinerators continue to find an important place in hazardous waste destruction. Variants on this include the use of cement kilns to burn organic materials.

Emissions of concern from incinerators are of a wide variety (▶ Hazardous Waste Incinerator Emissions). The formation of dioxins is certainly a prime topic. Dioxins are toxic chlorinated compounds likely formed from certain specific precursors containing chlorine. Their formation is strongly influenced in incinerators by the temperature. Dioxins partition to particulates in the stack and therefore can be somewhat removed along with the particulates in emission control features such as bag houses. Other POPs or persistent organic pollutants that constitute emission concerns include PCBs or polychlorinated biphenyls (which are also a dioxin precursors), PAHs or polycyclic aromatic hydrocarbons, PCNs or polychlorinated naphthalenes, HCB or hexachlorobenzene, chlorinated phenols, etc.

A prime emission of concern is the metal mercury. Other metals of concern include cadmium, lead, arsenic, and cobalt. Several mechanisms are being investigated to control these emissions.

Gaseous emissions include SO_2, NO_2, HCl, and HF. Incinerators are now often equipped with various flue gas–scrubbing devices which are effective at removing some of these gases, especially SO_2.

The ashes from incinerators also constitute a concern and are a hazardous waste in themselves (▶ Hazardous Waste Incineration Ashes and their Utilization). Newer technologies are in place to minimize the metal and POP content in ashes. Ashes are currently sometimes used in a variety of ways. One way that is frequent is the use of ashes to produce construction aggregates. Another way is to add to the mix in producing cements such as Portland cement.

There is a strong research interest in reducing and dealing with hazardous waste. Each year produces new and better technologies in this field.

Health Physics

JOHN W. POSTON, SR.
Department of Nuclear Engineering,
Texas A & M University, College Station, TX, USA

Article Outline

Glossary
Definition
Introduction
Early History and Development
The Health Physics Profession
Radiation Exposure to the US Population
Future Directions
Bibliography

Glossary

Absorbed dose The amount of energy deposited by ionizing radiation per unit mass of the material. Usually expressed in the special radiologic unit the rad or in the SI unit the gray (Gy). One Gy equals 1 J/kg or 100 rad.

Dosimeter Any device worn or carried by an individual into a radiation area to establish total exposure, absorbed dose, or equivalent dose (or the rates) in the area or to the individual worker while occupying the area.

Dosimetry The theory and application of principles and techniques associated with the measurement of ionizing radiation.

Epilation Loss of hair due to damage to the follicles in the skin. Temporary epilation occurs for acute exposures in the range of 300 to 500 rad with permanent epilation occurring above about 700 rad.

Equivalent dose (Formally the dose equivalent) The product of the absorbed dose and the radiation-weighting factor (formerly the quality factor) for the type of radiation for which the absorbed dose is measured or calculated. The equivalent dose is used to express the effects of radiation-absorbed dose from many types of ionizing radiation on a common scale. The special radiologic unit is the rem or in the SI unit the sievert (Sv). One Sv is equal to 1 J/kg or 100 rem.

Erythema Reddening of the skin due to exposure to radiation, similar to sunburn or thermal burns depending on the severity of exposure; occurs for acute exposures in the range 600–800 rad.

Exposure A quantity defined as the charge produced in air by photons interacting in a volume of air of known mass. An old quantity that is generally no longer used. Also, a general term used to indicate any situation in which an individual is being irradiated.

Ionization The process of removing one or more electrons from an atom or a molecule. The positively charged atom and the negatively charged electron are called an *ion pair*.

Isotope One of two or more atoms with the same number of protons (same atomic number Z) but a different number of neutrons (N) in their nuclei. A radioisotope is an isotope of a chemical element that is unstable and transforms by emission of nuclear particles and electromagnetic radiation to reach a more stable state. This term is often misused because unless the materials are the same element this term should not be used (see radionuclide below).

Nuclide A general term that indicates an atomic nucleus that is characterized by its atomic number (number of protons), number of neutrons, atomic mass, and energy state.

Radiation Used in this section to mean ionizing radiation. That is, particles or electromagnetic radiation emitted from the nucleus with sufficient energy to cause ionization of atoms and molecules, either directly or indirectly, composing the material with which the radiation is interacting.

Radionuclide A nuclide that is radioactive and upon decaying, emits ionizing radiation.

Definition

Health Physics: The theory and practice of radiation protection [1].

Introduction

The term health physics probably has a large number of definitions because it means many things to many practitioners of nuclear technology. As indicated above, health physics basically means radiation protection or radiation safety. The term itself has its origins in the Manhattan Engineering District (MED) days during World War II. Dr. Arthur Holly Compton and his deputy, Dr. Robert S. Stone, recognized the need for a group of individuals whose sole responsibilities were to protect the workers from the potential harmful effects of radiation while allowing the MED work to move forward as rapidly as possible. However, such individuals did not exist at the time and the decision was made to select individuals from the medical physics profession, who understood radiation exposures and biological effects of radiation, and those with expertise in measuring low levels of radiation exposure. The former group of individuals was recruited from the ranks of those in hospitals and cancer therapy clinics across the country while the latter group was recruited from cosmic-ray physicists (Compton himself was a cosmic-ray physicist). There was a need to select a name for these individuals and a number were considered. It was important that the name selected did not reveal the work of those involved in the MED activities. Since Dr. Stone was in charge of the Health Division at the Chicago Lab and the members of the group were physicists, the name health physics seemed to come naturally. There was also a clear intent to hide from those without the need to know exactly what was going on at the Lab. Thus, the term health physics was selected with the full intent that it be somewhat nebulous.

Dr. K. Z. Morgan, a cosmic-ray physicist and one of the original eight health physicists, defined health physics as the study and practice of radiation protection [2]. In a lighter moment, he also defined health physics as "what health physicist do." Health physicists are professionals dedicated to the protection of the human race and the environment from the harmful effects of radiation, while recognizing that there are enormous benefits to be derived from the use of nuclear technology [3, 4]. A health physicist is constantly balancing the risk of radiation exposure with the benefit from the use of the nuclear technology. In simple terms, health physicists strive to keep the radiation risk as small as possible and the benefit as large as possible. If one visualizes an old-fashioned balance or scale, the risk would be very small and the benefit would be large. It is not their intent to balance the scale but to have "benefit pan" outweigh the "risk pan" so that the scale is tilted significantly.

Current approaches to health physics (i.e., radiation protection) are based on three principles, summarized in three words: justification, optimization, and limitation. First, there should not be any unnecessary radiation exposures: the benefit must exceed the risk from the exposure. Second, all exposures to radiation should be kept As Low As is Reasonably Achievable (this is called the ALARA principle but, in other countries, this is called "optimization"). Finally, no radiation exposures should exceed the recommended regulatory limits on radiation exposures.

Those facilities licensed by either the US Nuclear Regulatory Commission (USNRC) or by agreement states generally apply an exposure limit of 5 rem/year (0.05 Sv/year) for occupationally exposed workers. But health physicists use the three principles to drive exposures as far below the regulatory limits as is possible. The ALARA principle has not only kept exposures below the regulatory limits, but also has resulted in a significant reduction in the collective doses and average worker doses at nuclear utilities in the United States. The average doses to workers at both pressurized water reactors (PWRs) and boiling water reactors (BWRs) in the United States have been reduced to levels of only a few hundred mrem per worker per year. In 2007, the average dose equivalent to a worker in the US nuclear power plant was 130 mrem [5].

Early History and Development

Even though the terms health physics and health physicist were not used until the early 1940s, the need to protect individuals from harmful radiation exposure dates back to just after the discovery of X-rays by Wilhelm Roentgen in 1885 and natural radioactivity by Henri Becquerel in 1896. With such amazing discoveries, many scientists began to investigate these phenomena for themselves and many were perhaps unnecessarily exposed to radiation. Many researchers took things into their own hands and established exposure guidelines and restrictions for use in their laboratories. However, often these exposure guidelines were based on observable effects such as the production of erythema and even evidence of epilation. Even with these informal and unified guidelines, there were still individuals who experienced harm from excessive radiation exposures.

In 1899, the British Roentgen Society formed a committee to collect data on the effects of X-rays. These efforts were hampered by the general belief that X-rays were harmless [6]. Later, the British X-ray and Radium Protection Committee adopted recommendations in 1921. It is important to notice the name of this committee because it indicates the limited nature of the early use of radioactivity and radiation-producing machines. In the United States, the American Roentgen Ray Society established a standing committee on radiation protection in 1920 and recommendations of this committee were adopted at the annual meeting of the Society in 1922 [2].

It was not until 1925, at the First International Congress on Radiology, that there was an international effort to begin to control radiation exposure. This effort resulted in the formation of the International Committee on X-ray and Radium Protection in 1928 at Second International Congress held in Stockholm [7]. These efforts resulted in two current organizations, the International Commission on Radiological Protection (ICRP) and the International Commission on Radiation Units and Measurements (ICRU). These international activities were paralleled by similar efforts in the United States. In 1929, the Advisory Committee on X-ray and Radium Protection was formed as an outgrowth of discussions between the American Roentgen Ray Society, the Radiological Society, and the Radium Society. This committee ultimately became the National Council on Radiation Protection and Measurements (NCRP) and was chartered by Congress in 1964 to provide advice and counsel on these matters to the federal government [7]. General overviews of the profession as well as discussion of the early development of the profession can be found in the bibliography provided at the end of this section [8–10].

The International Commission on Radiological Protection (ICRP) is responsible for making general recommendations on approaches to radiation protection. These recommendations are intended to provide the basic framework for radiation protection for those nations with nuclear technology programs, but it is the responsibility of each country to adapt these recommendations to the current state of their nuclear technology program. In the United States, the National Council on Radiation Protection and Measurements (NCRP) reviews the recommendations of the ICRP and provides a set of recommendations that are intended for use in the United States. The Environmental Protection Agency is responsible for establishing the general recommendations for all federal agencies and this guidance is signed into law by the President of the United States. Then, each federal agency is required to promulgate a set of regulations to govern their activities. For example, the USNRC regulations are found in Title 10 Code of Federal Regulations part 20 (10CFR20) and the Department of Energy regulations are in 10CFR835.

But, in the United States, things are not so simple. In the federal government, at least one agency has not revised their recommendations in response to the 1987 EPA guidance and is still functioning under the 1959 version of the Code of Federal Regulations. Most agencies are using the 1977 recommendations of the ICRP in compliance with the EPA guidance. The 35 agreement states in United States have promulgated regulations that are in compliance with the EPA guidance. However, the Department of Energy has instructed their contractors to implement the newer recommendations published by the ICRP in 1990. Even though these recommendations are similar, some significant differences must be recognized and understood. But, the situation in the United States is not currently very simple.

The Health Physics Profession

Health physicists are found in essentially all activities involving the use of ionizing radiation, including research, education, medical uses, industrial applications, consulting services, and state and federal regulatory agencies [3, 4]. Although most health physicists are involved in only one of these areas, it is common for the professional health physicist to become active in many of the above areas. There is always a need to train employees and educate the general public concerning the hazards of radiation and the appropriate methods of protection. A health physicist must ensure that the state and federal regulations regarding radiation exposure are followed and that legal limits are not exceeded. In addition, it is the health physicist's responsibility always to seek better ways of accomplishing tasks so that exposures to ionizing radiation are kept as low as reasonably achievable, releases to the environment are controlled and appropriate, and all uses of radiation or radioactivity result in a positive benefit.

Health physics is a course of study in a number of universities across the country; both at the baccalaureate and graduate levels. Many universities grant advanced degrees in health physics and related fields (physics, radiobiology, radiochemistry, etc.). The courses of study are found in colleges of engineering, schools of public health, physics departments, biology departments, and chemistry departments. There seems to be no official "home" for these programs. However, this situation provides a certain interdisciplinary flavor to the profession, because each academic program is different and students can find an institution that best satisfies their interests and needs.

Health physics is an exciting, interesting, and challenging profession. Controlling radiation exposures to the lowest possible levels, while allowing the use of nuclear technology for the benefit of society, requires a broad background in a number of technical areas. This broad general understanding includes physics, chemistry, biology, anatomy, physiology, ecology, electronics, and a number of other disciplines. Opportunities for employment exist within a number of allied professional activities, such as medicine, cancer therapy, research, education, and regulatory affairs. Health physics and its related fields present the individual with daily challenges that require the application of sound professional judgment so that other work can proceed safely.

Opportunities within the profession have increased dramatically over the past few years. Projections indicate that the demand for health physicists will exceed to supply for the foreseeable future. Many of the early leaders of the profession have retired and a large number of those active in the profession have more than 30 years of experience. The demands for qualified health physicists will likely stay high for a very long time as the use of nuclear technology continues to increase.

Radiation Exposure to the US Population

Health physicists have the responsibility to protect the general population from unnecessary radiation exposure, just as they do with occupationally exposed workers. A question that often arises is, "what if the average radiation exposure to members of the public?" The NCRP has answered this question in a series of reports over a number of years. In 1987, the NCRP estimated that the average dose equivalent to a member of the US population was 360 mrem/year [11]. A large majority of this exposure was due to naturally occurring radiation sources such as cosmic rays, solar radiation, internally deposited radionuclides, and, especially, radon gas (200 mrem/year). In the most recent NCRP report [12], the average equivalent dose was estimated to be 620 mrem/year (6.2 mSv/year). The significant increase in this average was due to the use of medical technology involving radioactivity (nuclear medicine) and radiation-producing devices used in diagnosis and in cancer therapy (48%). Occupational exposures to radiation were estimated to be <0.1% of the total average equivalent dose. This is perhaps a testament to the dedication of health physicists to keep these exposures ALARA.

Future Directions

The practice of radiation safety and control (health physics) will continue to evolve as more and more challenges are added to the daily routine of the professionals. Revisions of dose/risk estimates continue to result in lower recommended radiation exposures limits. Some of these may be unjustified but, nevertheless, place an increased burden on those responsible for occupational exposures. One big challenge in the United States is the impact of the use of radiation-producing machines and radioactive materials in medicine. The increase in the average annual exposure to the general population is astounding and is a concern to those dedicated to the profession of health physics and the concept that all exposures should be justified.

Bibliography

1. Poston JW Sr (1987) Health physics. In: Encyclopedia of physics science and technology, vol 6. Academic, New York
2. Morgan KZ, Turner JE (1973) Principles of radiation protection. Robert E. Kreiger, Huntington
3. Health Physics Society, Health physics: a profession of the nuclear age, Health Physics Society, 1313 Dolley Madison Blvd., Suite 402, McLean, VA 22101
4. Health Physics Society, Prospectus, Health Physics Society, 1313 Dolley Madison Blvd., Suite 402, McLean, VA 22101

5. Nuclear Energy Institute (2009) Radiation safety at nuclear power plants, June 2009. http://www.nei.org/keyissues/safetyandsecurity/factsheets/safetystudiespublicworkers. Accessed 20 May 2010
6. Kathren RL, Tarr NE (1974) The origins of the Health Physics Society. Health Phys 27:419–428
7. Taylor LS (1971) Radiation Protection Standards. CRC, Cleveland
8. Taylor LS (1958) Brief history of the National Committee on Radiation Protections and Measurements (NCRP) covering the period 1929–1946. Health Phys 1:3–10
9. Taylor LS (1958) History of the International Commission on Radiological Protections (ICRP). Health Phys 1:97–104
10. Claus WD (1958) What is health physics? Health Phys 1:56–61
11. National Council on Radiation Protection and Measurements (1987) Ionizing radiation exposure of the population of the United States. National Council on Radiation Protection and Measurements, Bethesda
12. National Council on Radiation Protection and Measurements (2009) Ionizing radiation exposure of the population of the United States. National Council on Radiation Protection and Measurements, Bethesda, NCRP Report 160

Heat Content and Temperature of the Ocean

Rui M. Ponte
Atmospheric and Environmental Research, Inc,
Lexington, MA, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Brief History of Temperature Observations
From Raw Observations to Scientific Analyses
Spatial and Temporal Patterns of Variability on Climate Scales
Future Directions
Acknowledgments
Bibliography

Glossary

CTD (conductivity-temperature-depth) Instrument used to obtain vertical profiles of salinity, temperature, and pressure in the ocean at resolutions up to 1 m.

Heat content Amount of energy contained in thermal motion associated with any volume of matter, which is proportional to its temperature and specific heat.

Mixed layer Near-surface layer where turbulent mixing primarily induced by atmospheric forcing produces homogeneous conditions of temperature, salinity, and other water properties.

Objective analysis General mathematical procedure by which the value of a variable of interest at a given location and time is derived from similar observations at other times and/or places (also sometimes referred to as optimal interpolation).

Potential temperature Temperature attained by a fluid parcel if measured at a given standard pressure (typically 1,000 hPa corresponding approximately to atmospheric pressure at the ocean surface).

Specific heat Amount of energy per unit mass necessary to change the temperature of a given substance by 1°C, with typical values for seawater around 4,000 J/kg/°C.

Thermistor Sensor that uses known thermal dependences of electrical resistance to determine temperature.

Thermocline Region of enhanced vertical temperature gradients separating well-mixed layers near the surface from the layers with weak thermal stratification found in the deep ocean.

XBT (eXpendable BathyThermograph) Instrument used to obtain essentially continuous recordings of temperature with depth, from the surface to about 700 m, which can be deployed from a ship underway and without having to be retrieved.

Definition of the Subject

The global ocean is a vast body of water covering approximately two thirds of Earth's surface and having an average depth around 4,000 m. Temperatures in the ocean can range from a balmy 30°C in the surface layers of the tropics to an icy cold −2°C in polar regions and at abyssal depths. Apart from the strong spatial variability, water temperatures at a place also change on all time scales, from hours to decades. Variability in surface water temperatures affects air-sea heat exchange and is an important factor determining the nature and strength of ocean–atmosphere coupling. Changes in ocean temperatures imply expansion or contraction of the water column and can thus have a major impact on sea level. And given the large heat capacity of seawater compared to that of air and the large mass of the oceans compared to that of the atmosphere, relatively small fluctuations in oceanic temperatures imply substantial changes in heat content, which are essential for properly accounting for the planet's energy balance. As such, taking the temperature of the oceans and determining changes in its heat content is an essential diagnostic of the state of Earth's climate and the overall health of our planet.

The task of monitoring seawater temperatures is made very difficult by the vastness of the global oceans and by the remote and harsh environment in which measurements must be taken. Original temperature observations date back at least a few centuries and were made by dropping thermometers in the water, first at the surface and then also at depth. Developments mainly in the second half of the twentieth century have led to a number of different sensor technologies, including continuous vertical profiling instruments deployed on various platforms (ships, moored instruments, autonomous floats), acoustic tomography, as well as space remote sensing techniques. Despite significantly improved data gathering capabilities, considerable challenges remain in observing the oceans globally and at sufficient temporal and spatial resolution. The sampling problem is a daunting one, and the historical temperature database is fraught with other issues as well regarding lack of metadata, unknown instrumental biases, and consistency between shifting technologies. A number of developing observing systems, together with advancements in modeling and data synthesis capabilities, promise to deliver a more accurate chart of ocean temperature and heat content in the years ahead.

Introduction

Measurements of seawater temperatures must have started with the appearance of the first thermometers in the seventeenth century. As early as the eighteenth century, Benjamin Franklin was making use of a thermometer as a rudimentary navigation aid to chart the changes in temperature across the Gulf Stream [1]. With the development of commercial shipping and the birth of well-funded military naval fleets, routine ship measurements of near-surface temperature were well under way in the nineteenth century. Dedicated scientific efforts to sample the oceanic temperature structure, not just at the surface but also at depth, had to wait until the second half of the nineteenth century, however. The global Challenger expeditions in the 1870s [2] and later the Meteor expeditions in the 1920s [3] signaled the birth of oceanography as a science. By this time, the importance of knowing the physical and chemical properties of seawater had become all too clear.

Temperature is a key parameter for determining water density, along with salinity and pressure. As it warms and cools, seawater expands and contracts by an amount determined by its coefficient of thermal expansion, which increases with temperature, pressure, and salinity (e.g., [4]) and can vary significantly from place to place in the ocean. Density changes tend to be more sensitive to temperature at depth, where pressure is largest, and in the tropics, where the waters are warmest. As horizontal differences in density determine pressure gradients and currents, temperature is of crucial dynamical importance when trying to infer the ocean circulation and how it changes in time.

The impact of temperature on density (or correspondingly volume) also means that changes in ocean thermal properties are directly related to changes in sea level. As described in most basic oceanographic texts (e.g., [4]), by knowing the fluctuations in temperature throughout the water column, their integrated effects on volume can be calculated and represent the so-called thermosteric component of sea level variability. (Salinity also affects volume and gives rise to halosteric changes in sea level. Steric sea level signals refer to the combined effects of temperature and salinity.) Interpretation, analysis, and prediction of sea level change are thus tightly related to accurate knowledge of temperatures over all ocean depths.

Last but not least, seawater temperatures give a measure of how much thermal energy or heat content is contained in the ocean. Oceanic heat content is a quantity of major relevance for climate and is a particular focus of this entry. Given the difference in the specific heat of water and air, heating or cooling a given amount of seawater requires approximately four times more energy than for the same mass of air.

More importantly, because the density of water is approximately 1,000 times larger than the density of air, the ocean's heat capacity and thermal inertia are vastly larger than those of the atmosphere. As such, the oceans can serve as a large reservoir of heat in the climate system. Being a dynamic fluid with an active circulation, the oceans continually transport and redistribute heat and participate in important heat exchanges with the atmosphere as well as the cryosphere. Knowledge of oceanic temperature fields can thus hold vital information on the past, present, and future state of Earth's climate.

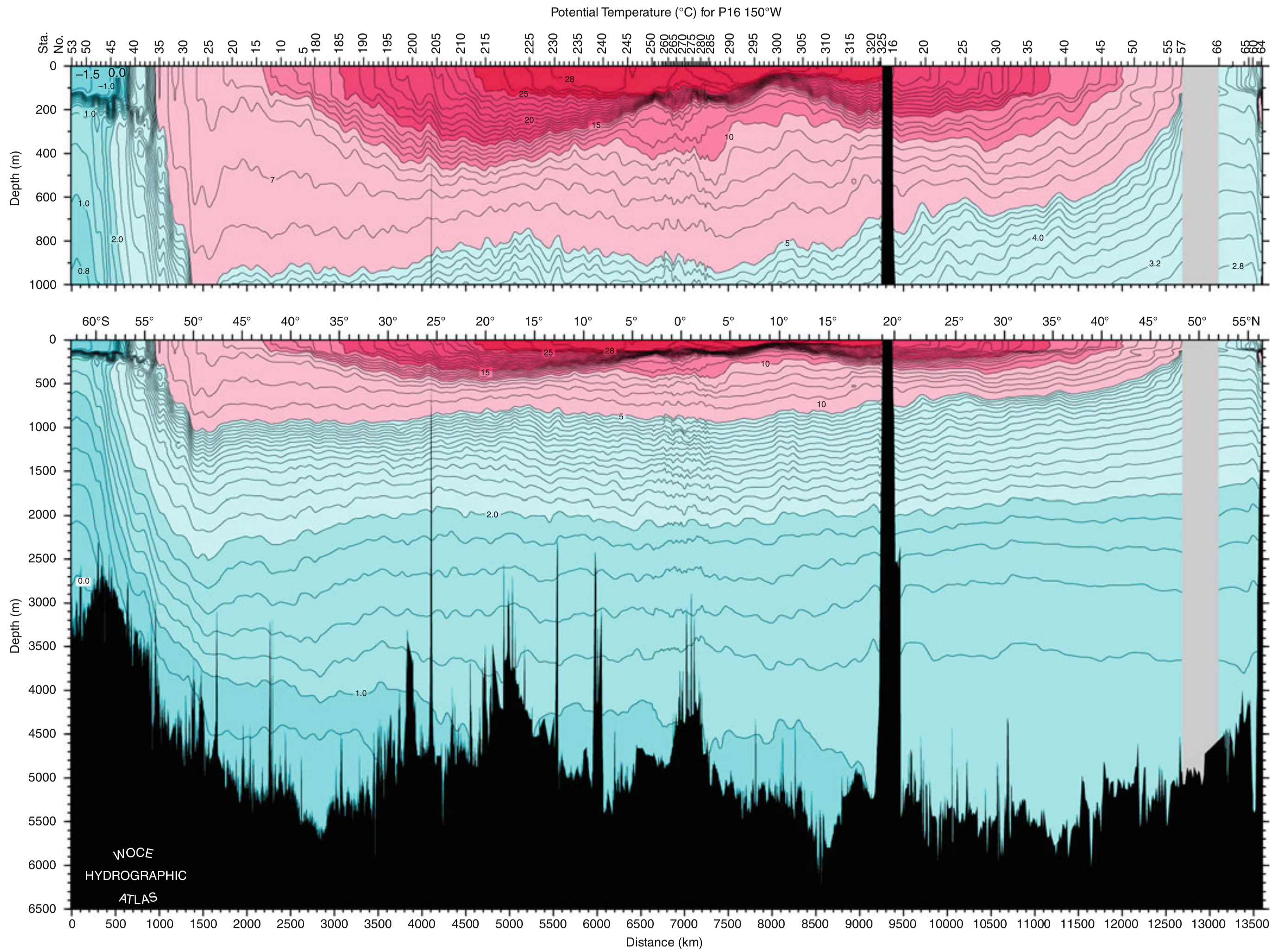
By the time the classic oceanography text by Sverdrup et al. [5] was published, summarizing much of what was known about the oceans at the time, the crude characteristics of the oceanic temperature structure were beginning to take shape. Apart from the clear decrease of near-surface temperature with latitude away from the tropics, mimicking that of surface air temperatures, other major features, well known to us today and seen in the example of Fig. 1, were already apparent. The presence of a near-surface layer with homogenized temperature and other properties was established early on. This so-called surface mixed layer owed its existence to high turbulence and mixing induced primarily by air-sea fluxes at the upper boundary, and its thickness, which could reach several hundred meters, was very much dependent on location. Below the mixed layer, a region of rapid decrease in temperature with depth was present in most of the oceans. The rapid transition between the warm surface layers and the cold abyss, nominally called the thermocline, exhibited a fairly complex dependence on region and could extend over various thicknesses and up to depths of 1,000 m. Much colder deep and bottom waters beneath the thermocline tended to be less stratified with depth and also more homogeneous in latitude and longitude, without the marked latitudinal dependence of the upper layers.

The initial picture that emerged from the first scientific expeditions in the nineteenth and twentieth centuries was necessarily one of essentially steady conditions, given the paucity of data with a few snapshots taken along a handful of sections. Ever since those original efforts, the quest for determining temperature distributions in space and time over the vast global oceans has been an endeavor with almost epic overtones. In what follows, a brief description of the panoply of instruments and methods used to measure temperature and heat content is attempted within a historical timeline, from the earliest to the most modern period. Such a voyage through the world of instrumentation and data gathering sets the stage for a discussion of what is presently known about ocean temperature and heat content, starting with the near-surface layers and progressing down to the abyssal depths. The main focus is on the thermal structure at large spatial scales (100 km and longer) and its variability on time scales of months to centuries relevant for climate. Although much progress has been achieved in recent decades, considerable challenges remain and some possible pathways for the future are traced at the end.

Brief History of Temperature Observations

Conventional thermometers were the first instruments used to sample oceanic temperatures [5]. The bulk of the data was taken at the surface, either by dropping the thermometers in the water or by using a bucket to collect water for measurement. Some subsurface measurements were obtained by using insulated bottles that would maintain collected water samples at their ambient temperatures, with minimal exchanges of heat with surroundings during ascent. For subsurface sampling, however, most scientific surveys through the first half of the twentieth century were made using reversible thermometers [5], which were set to sample at some predetermined depth levels. The technology was simple and reliable but expensive. Furthermore, only point measurements in time and space were possible with reversible thermometers.

Continuous sampling devices were experimented with early on but their use did not become widespread until the second half of the twentieth century [7]. First thermographs used standard thermometry and various methods for continuous data recording and were mainly deployed in the hulls or intake pipes of ships. Such instruments had no pressure sensors and thus no ability to determine depth, and data gathering was confined mainly to the top few meters of the ocean. The first useful attempts to continuously record temperature and depth are reported by Spilhaus [8, 9]. The original bathythermograph (BT) followed a completely



Heat Content and Temperature of the Ocean. Figure 1
Latitude-depth cross-section of potential temperature along approximately 150°W in the Pacific. *Contour lines* are drawn every 1°C (0.2°C) for temperatures warmer (colder) than 5°C. Different *color shading* is shown at intervals of 5°C, with temperatures getting progressively colder from *red* to *blue shades*, and with the *darkest red* representing the range 25–30°C. *Bottom panel* shows the full water column, with the *top panel* focusing on the *upper* 1,000 m for added detail. Bathymetry is shaded in *black*, with the land feature reaching the surface near 20°N being the Hawaiian Islands (Reproduced from Talley [6])

mechanical design and its reliability made it one of the most extensively used instruments for ocean surveying in the 1950s and 1960s [10]. Its depth range was quite limited, however, allowing measurements only in approximately the upper 150 m.

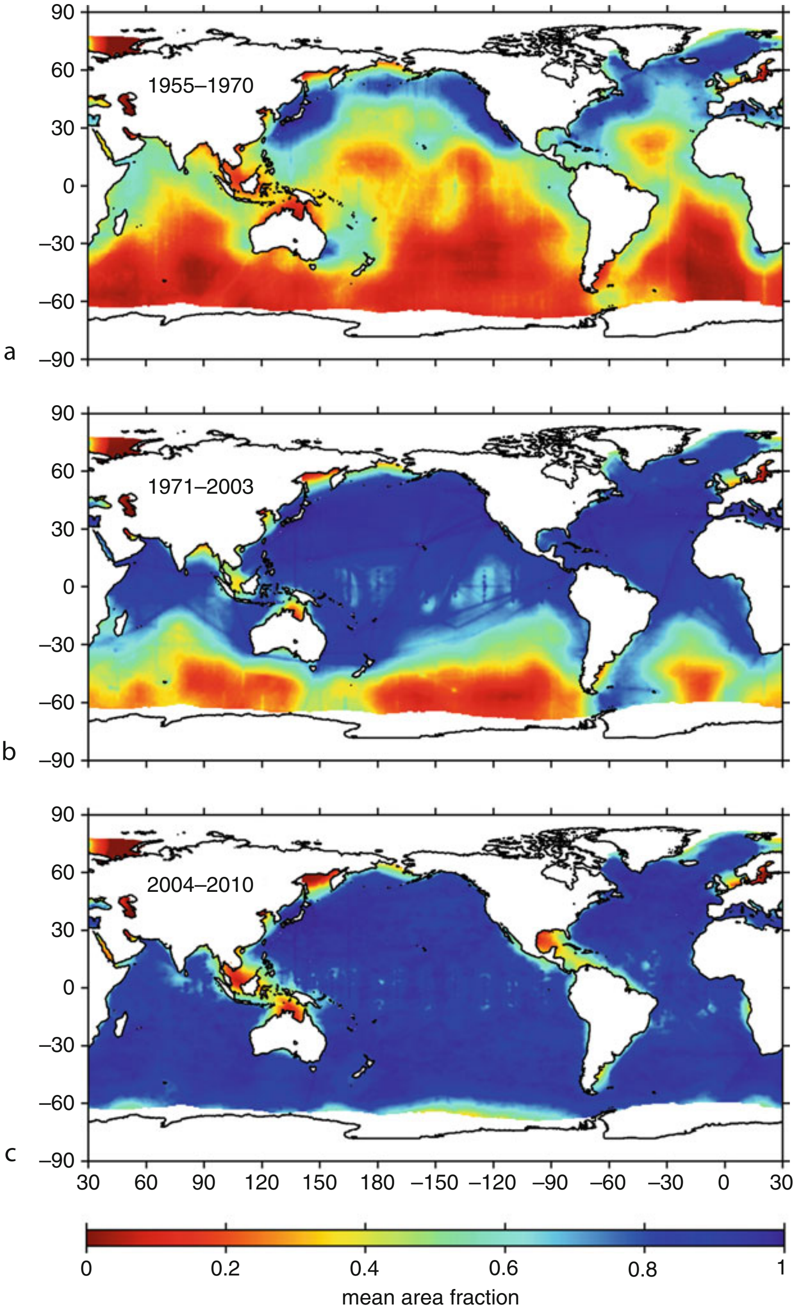
An important development was the appearance of the expendable BT (XBT), which used thermistor technology to measure temperature over approximately the upper 700 m of the ocean [10]. Without a pressure sensor, the depth of measurement was inferred from the elapsed time since deployment and the "known" free fall rates, as established by the manufacturer of the instruments. Because of its relatively low cost and its easy use and operation from a ship underway, the XBT became the workhorse of the in situ observing system for a long time. A great deal of what is known about the upper ocean temperature and heat content variability is owed to millions of XBT profiles carried out mostly from merchant ships and other ship-of-opportunity vessels. As such, XBT sampling tended to be best along most traveled shipping routes, particularly between Europe and North America (see Fig. 2). By any measure, large chunks of ocean were rarely visited by XBT surveys and the deep ocean was totally out of reach. The method of assigning depths to XBT temperature profiles based on instrument free fall rates also led to the introduction of complicated biases, which are still being dealt with to date [12, 13].

In parallel with XBT development, the need for high-accuracy profiling instruments that could sample the ocean all the way to the bottom led to the development of the conductivity-temperature-depth (CTD) sensor. Various different versions were produced mainly in the 1950s and 1960s, and CTD use quickly became the standard in all the major scientific campaigns (e.g., [7]). Compared to XBT deployments, which could be carried out with the ship underway, CTD casts required dedicated ship time, but CTD precisions were significantly better than those possible with XBT sensors, and including pressure sensors ensured an accurate mapping of the thermal structure in the vertical. The CTD instrument has continued to evolve and current versions have accuracies better than 0.001°C and vertical resolutions of about 1 m. (Temperature variability at submeter scales is a topic in its own right and much of the work that led to the CTD sensor has evolved into several instruments capable of measuring thermal microstructure at millimeter scales; interested readers can learn more about microstructure instrumentation by consulting for example [7, 14, 15] and their cited references.)

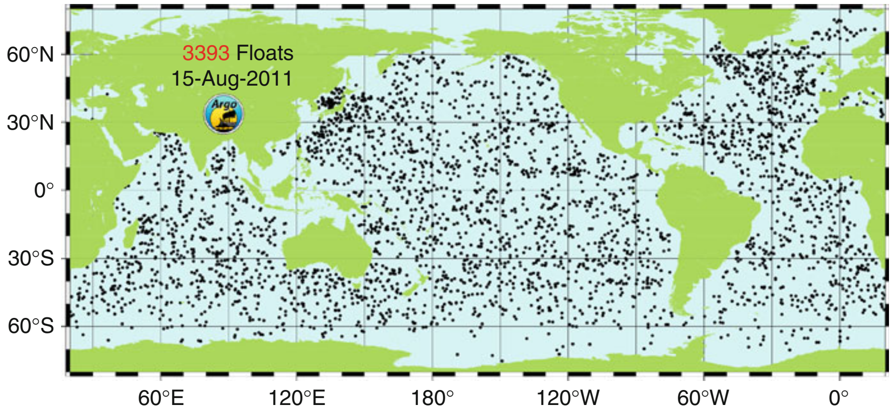
Traditional CTD surveys, involving basin-wide sections of full-depth casts made at variable horizontal spacings (typically a few tens of kilometers, as in the example of Fig. 1), together with XBT profiles along major shipping lanes, continued to be the main method of observing oceanic thermal structures until the end of the twentieth century. But even as the World Ocean Circulation Experiment (WOCE) became a reality [16], focus was still on obtaining a single realization of the global temperature field. Only a few sections were meant to be sampled repeatedly to assess the full temporal variability. Costs of operating research ships from which CTD casts could be performed continued to be a major impediment for truly global coverage. Even after several decades of XBT and CTD measurement campaigns, extensive oceanic regions, particularly in the Southern Hemisphere, remained severely under-sampled by the end of the twentieth century (Fig. 2).

The march toward a truly global observing system took another major step with the implementation of the Argo system of profiling floats in the last decade [17, 18]. Following years of research and development in autonomous float technology, major international cooperation under the Argo Project led to the seeding of the global ocean with more than 3,000 floats (Fig. 3) with the capability of profiling temperature and other properties between depths of 2,000 m and the surface approximately every 10 days. Measurements are relayed by satellite while the floats are at the surface in between each cycle of ascent and descent. Coverage is almost global, apart from the lack of data in high latitudes and ice-covered regions, where current Argo technology does not work (Fig. 3). Shallow and marginal seas are also not part of the sampling Argo strategy. In a short number of years since their deployment, Argo floats have produced more than 300,000 profiles and have provided more samples than all the data ever collected before in many regions.

Despite all the efforts to measure subsurface temperature, it is at the surface that coverage is most adequate in both space and time (e.g., [19]). A major factor in the density of observations at the surface has



Heat Content and Temperature of the Ocean. Figure 2
Estimate of the annual mean sampling for the *upper* 750 m expressed as the fractional area of data coverage, provided by John Lyman as an update to the results discussed in Lyman and Johnson [11]. (**a**) Coverage prior to 1970 is very poor over most of the global oceans. (**b**) Substantial improvements in the period 1971–2003 come from XBT surveys, mostly in the Northern Hemisphere and along ship routes clearly seen as tracks in the plot. (**c**) The increase in coverage for the most recent period shown (2004–2010) reflects data collected by the Argo system



Heat Content and Temperature of the Ocean. Figure 3
Distribution of Argo floats as of August 2011 as estimated by the Argo Project (Reproduced from the Argo Project web site). Global coverage of the oceans involves more than 3,000 floats. High northern and southern latitudes and shallow and marginal seas remain poorly sampled

been the space-based remote sensing techniques established since the 1980s (e.g., [20, 21]). Given the dependence of surface infrared and microwave radiation on sea surface temperature, satellite radiometers measuring at these frequency bands can be used to retrieve sea surface temperature. Many other factors like atmospheric clouds and water vapor impact the radiation reaching the instrument and have to be corrected. For these purposes, increasingly complex validation and calibration methods have been developed over the last 30 years. Many of such efforts rely on in situ measurements from ships, surface buoys, and drifters. However dense in space and time, satellite measurements strictly measure the temperatures right at the surface and their interpretation in terms of near-surface temperatures remains an issue of research [21].

For subsurface thermal information, aside from profiling measurements by XBT, CTD, and Argo instruments, a number of other in situ programs are also useful to mention. Surface and subsurface moorings of various types have been deployed, which usually include thermistor chains. Operational programs are in place in the tropical Pacific and Atlantic [22, 23] with a similar system currently under implementation in the tropical Indian Ocean [24]. Many research projects with local or regional focus provide occasional temperature measurements (e.g., [16, 25]), and a few long-term stations have also been maintained in places like Bermuda and Hawaii [26, 27]. In any case, traditional ship-based XBT and CTD surveys and more recently the Argo floats provide the bulk of information on ocean temperature and heat content over the last decades.

Apart from direct temperature measurements, several other indirect observing methods have been recently developed. Indirect thermometry can make use of the relationship between temperature and any other oceanic variable that can be accurately measured. Such is the case with the strong dependence of sea level on thermosteric height, which is directly related to integrated temperature over the water column. Measuring variability in sea level using satellite altimetry can complement in situ temperature observations in efforts to infer changes in heat content [28, 29]. As with many indirect methods, ancillary observations are needed. As sea level also depends on salinity through changes in halosteric height, measurements of salt content can help in isolating the temperature effects. In addition, changes in sea level can be due to changes in the mass of the water column and the ability to measure the total mass content either using in situ bottom pressure recorders or satellite gravity techniques can lead to improved inferences of heat content based on sea level observations, as discussed by Ponte [30] and Jayne et al. [31].

Temperature has also a strong effect on sound speed. Such dependence is the basis for the acoustic thermometry ideas proposed by Munk and Forbes [32] and others. Sound travel times can be inverted to estimate variability in averaged temperature and heat content along ray paths connecting any source and receiver pair. Plans to establish acoustic monitoring for the global oceans are revisited in Dushaw et al. [33] but presently current data sets are confined to a few experiments carried out mostly in the North Pacific, between Hawaii and the North American coast [34]. Long-range acoustic thermometry can work over many thousands of kilometers and thus can average noisy mesoscale fields over large oceanic volumes at the expense of lack of horizontal resolution of any temperature or heat content features. Although the vertical resolution is limited, acoustic measurements can provide much needed temperature estimates at depths presently not sampled at all by the Argo system.

From Raw Observations to Scientific Analyses

From a pure observational perspective, the diversity of instruments and observing platforms involved in temperature data gathering at any one time, the shifts in technology throughout many decades of observations, the lack of detailed information about many historical data sets (e.g., types of XBT instruments used, depth at which near-surface temperatures are taken), the various parties involved in collecting observations, among other factors, make archiving and quality controlling the data a laborious and difficult task. Major efforts in the USA have been pursued at NOAA's National Oceanographic Data Center and are well documented in the World Ocean Database series produced by NOAA roughly every 4 years, with the last edition released in 2009 (e.g., [35, 36]). Other more regional compilations are also available (e.g., [37]), and no central archive includes all the collected temperature data. Data processing tasks commonly involve statistical checks for outliers, elimination of duplicate observations, correcting for depth biases, along with many subjective criteria for flagging "suspicious" observations. More restricted data compilations based on specific hydrographic programs like WOCE [38] allow for better quality control and accuracy. In the end, researchers must decide what data to use and to discard (e.g., [38, 39]) and scientific analysis and results can depend on such choices (e.g., [37, 40, 41]).

Apart from data quality issues, the irregular and sparse sampling in both space and time makes quantitative study of temperature and heat content variability for climate inference far from a trivial task. Most studies rely on objective analysis/optimal interpolation methods (e.g., [42, 43]) to fill in data gaps in both space and time and reduce the observations to regularly gridded fields in latitude and longitude at given standard depth levels, which are more prone to scientific analysis. Objective analysis methods require assumptions about the nature of the fields being estimated (e.g., their degree of smoothness in space) and the statistics of the data (e.g., spatial and temporal correlation scales of data noise), which can influence how each data point is weighted to provide an estimated value at places where no data are available. As with data use, subjective choices are often involved and can lead to differences in results, particularly for regions where data are very sparse and estimates become more dependent on the details of the objective analysis method used. The series of World Ocean Atlas from NOAA (e.g., [44]) are a widely used example of a mean and seasonal temperature climatology produced from objective analysis applied to the wide variety of temperature measurements contained in the World Ocean Database [35, 36].

More advanced methods of data interpolation involve the use of ocean general circulation models (GCM), which embody the most complete description of the dynamic relations among all relevant observables (temperature, salinity, velocity, sea level) under strict laws of fluid motion and conservation principles of heat, momentum, and other quantities. Various data assimilation or state estimation techniques (e.g., [43, 45]) allow optimal combination of the information embedded in GCM codes with that contained in observations of all types. The use of such computationally intensive methods for climate analyses has been common practice in meteorology (e.g., [46]) and is becoming more common in oceanography (e.g., [47–49]) after rapid developments in ocean GCMs and data gathering capabilities, as well as in computing power, in the last decade. Most techniques involve minimization of an "objective" function that evaluates model-data misfits and thus require knowledge about data error statistics, which determine how tightly the model is allowed to fit the data. Books by Daley [43], Wunsch [45], and several others discuss the various basic approaches to data assimilation available. For the purposes of analyzing oceanic temperature and heat content variability in a climate context, it is helpful to use methods that produce temperature evolutions consistent with all the flow fields, air-sea fluxes, and other boundary and initial conditions. As discussed by Wunsch et al. [48], fitting the temperature observations must be achieved if possible through dynamically consistent adjustments in currents, atmospheric forcing, model physics, and other factors, and without introducing unphysical jumps in the temperature evolution, which would imply the violation of basic heat conservation principles.

Estimates of temperature fields, obtained either through objective analysis or model-data synthesis methods of the sort described above, provide the basis for a variety of climate analyses, including studies of air-sea interaction, sea level change, planetary heat balance, and many other topics. In the context of this entry, the focus is on oceanic heat content variability. Changes in heat content per unit area are just proportional to changes in temperature ΔT over a layer of thickness H and given by $\rho C_p H \Delta T$, where ρ is the density of seawater and C_p is its specific heat at constant pressure [4]. For typical values of $\rho \sim 1{,}028\ \mathrm{kg/m^3}$ and $C_p \sim 4{,}000$ J/kg/°C, increasing the mean temperature of the water column (~4,000 m) by 1°C implies a change in heat content of around $1.6 \times 10^{10}\ \mathrm{J/m^2}$. To achieve such warming over 1 year, an anomalous heat flux of $\sim 500\ \mathrm{W/m^2}$ would have to be applied. By the same token, an error in the annual change in mean temperature of 0.001°C amounts to an error in implied heat flux of $0.5\ \mathrm{W/m^2}$. For comparison, present estimates of the planetary heat flux imbalance under possible global warming scenarios are of the order of $1\ \mathrm{W/m^2}$ [50]. Thus, very small changes in oceanic temperatures can imply large changes in heat content, in terms of Earth's heat balance and climate, and accurate global measurements of ocean temperatures, from the surface to the bottom, and from the tropics to the polar regions, are deemed essential in this context.

Spatial and Temporal Patterns of Variability on Climate Scales

A vast literature documents numerous attempts at describing the temperature and heat content variability in the global oceans, either from straightforward analysis of the temperature observations (e.g., [51]), from objective analysis methods using only temperature or also other types of data (e.g., [29]), or from data-constrained model solutions based on the most sophisticated assimilation methods (e.g., [52]). In this section, a summary of the current state of knowledge on the subject is presented, focusing on the advances made in most recent years. Information on the upper ocean layers is considerably richer in detail than on the deeper layers, as the amount of data quickly dwindles with increasing depth. Temperature and heat content analyses are discussed interchangeably. The simple relation between temperature and heat content discussed in the previous section can be used to approximately convert between the two variables.

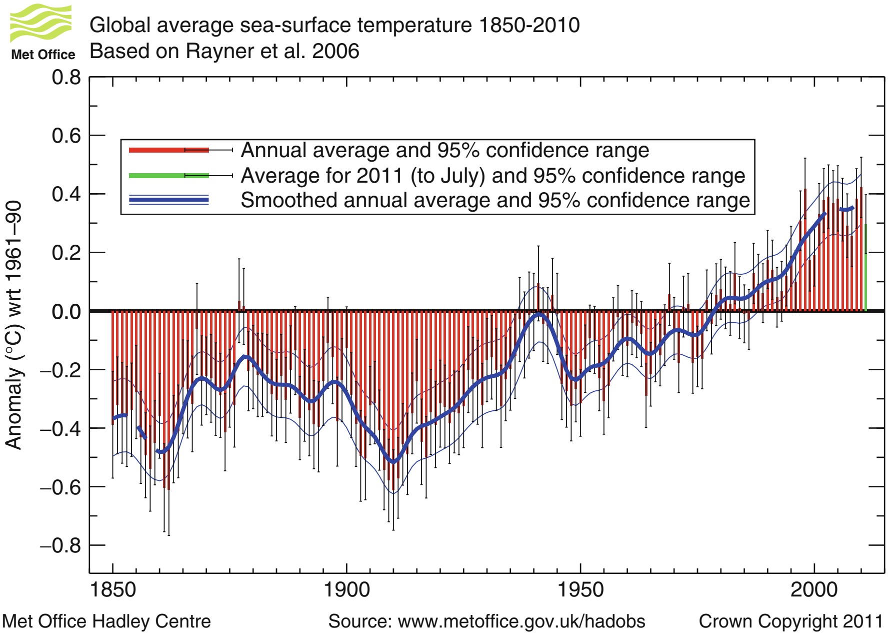
Spatial and temporal coverage of the near-surface ocean is far better than at any other depth level, given the early practices of surface measurements from ships and more recently the global daily coverage brought by the advent of the satellite era. These widely different measurement techniques have been combined to produce records of what is commonly called sea surface temperature (SST). What SST values represent is, however, not a trivial issue because of the presence of vertical temperature gradients very near the surface. Much of the in situ data comes without precise information on the depth of measurement, which can be anywhere from near the surface to a few meters below it. In turn, satellite sensors sample the upper few microns or millimeters, depending on whether infrared or microwave instruments are used. As such, satellite values can be influenced by strong diurnal effects affecting the skin layer and can also differ substantially from "bulk" SST properties representing an average of the upper few meters (~10 m) of the ocean. A detailed account of many of these issues, including various definitions of SST and how in situ and satellite data are being used to improve present and past SST records, can be found in Donlon et al. [21] and references therein.

Despite all the inherent ambiguities, SST records constitute one of the most important oceanic datasets for climate studies. Historical data compilations (e.g., [19]) provide estimates of SST since the middle of the nineteenth century, although the earlier periods have much sparser data coverage (e.g., [53]).

When global mean estimates are attempted, a record of increasing SST over the last century emerges from most studies, consistent with a warming planet [20, 54, 55]. As can be seen in Fig. 4, present global mean SST estimates are almost 1°C warmer than in the early twentieth century. Apart from a long-term warming trend, variability at all time scales is present in the annual mean values. Considerable uncertainties are involved, however. Time-varying instrumental biases are a particular concern (e.g., [56]) and efforts to correct such errors are likely to continue (e.g., [57]).

While global mean SST is an essential climate index, just like surface air temperature, numerous climate variables and processes are related to SST spatial variations, from the intensity and quantity of tropical cyclones, which depend on tropical warm waters to form and grow, to the onset of deep convection in the ocean, which can happen in the presence of cold, dense surface waters at high latitudes. A great deal of effort has thus been devoted to understanding major spatial and temporal patterns of SST variability in relation to air-sea fluxes, ocean and atmosphere dynamics, and their connection to climate variability. Deser et al. [53] review major modes of interannual and multi-decadal variability with clear SST signatures, including the El Niño-Southern Oscillation, the Pacific Decadal Oscillation, the North Atlantic Oscillation, and the Atlantic Multi-Decadal Oscillation. As one parameter controlling air-sea transfers, knowledge of SST is key to making long-range climate forecasts (e.g., [58]). In addition, the existence of century-long time series of SST with some spatial information provides essential surface boundary conditions for atmospheric model simulations of the past climate and allows for model calibration and validation (e.g., [46, 59]).

Depending on region and time of year, SST observations can provide a window on mixed layer temperatures and can thus yield information on heat content in the near-surface layers of the ocean, assuming the mixed layer depth is known. Ultimately, for a full and quantitative description of heat content variability, subsurface temperature observations are needed. Until the implementation of the Argo system, a great majority of temperature profiles were based on XBT casts and confined to the upper 700 m. Multidecadal analyses of large-scale temperature and heat



Global average sea-surface temperature 1850-2010
Based on Rayner et al. 2006

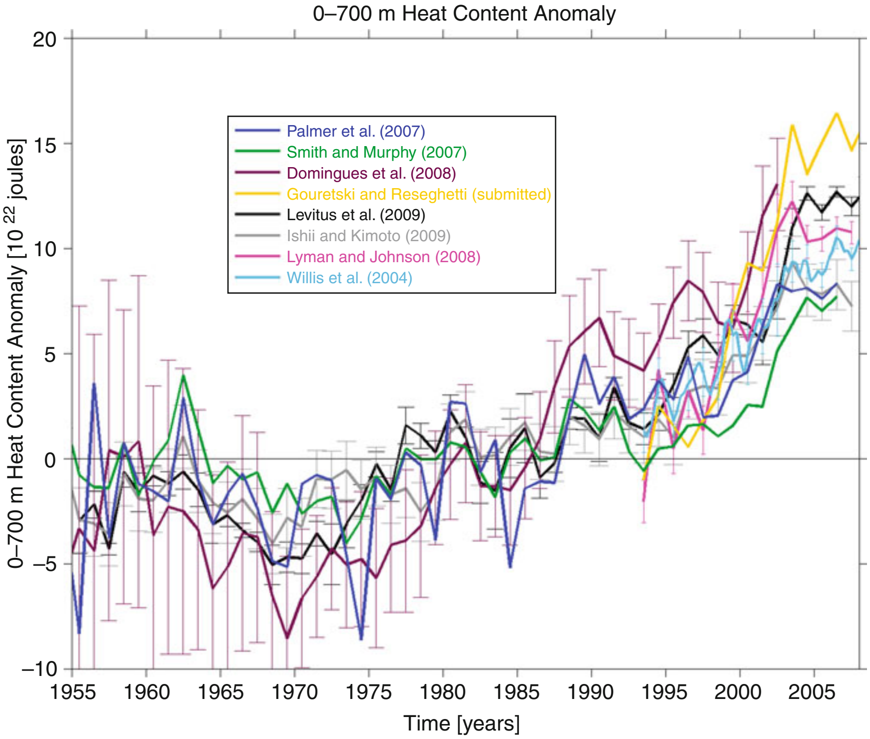
Met Office Hadley Centre Source: www.metoffice.gov.uk/hadobs Crown Copyright 2011

Heat Content and Temperature of the Ocean. Figure 4

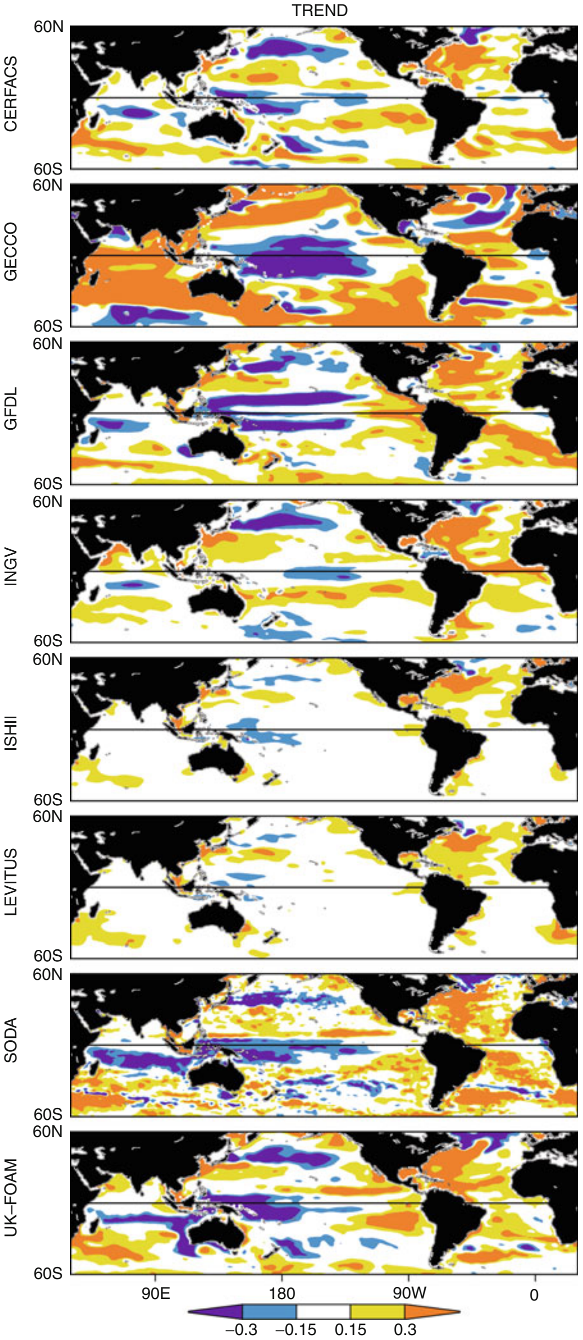
Global mean sea surface temperature for the period 1850–2009 (anomalies relative to 1961–1990) based on Rayner et al. [54] and produced by the UK Met Office Hadley Center (http://www.metoffice.gov.uk/hadobs) content variability have thus focused on these depths. Early observations are, however, quite thin and time series studies are restricted to the period since the 1950s. Apart from the presence of a strong seasonal cycle, which is mainly confined to the upper 100–200 m of the water column (e.g., [60, 61]) and driven by heat exchange with the atmosphere, the main interest has been on determining variability in temperature and heat content at scales of years to decades.

Numerous temperature analyses have produced estimates of mean oceanic heat content anomalies for depths 0–700 m (e.g., [37, 41, 62]). Several such time series, all using mostly the same available data but different processing procedures, are reproduced in Fig. 5 [63]. Most analyses indicate warming and increasing heat content of the upper ocean over the last three decades, in agreement with analyses of SST (see Fig. 4). The long-term warming has been related to anthropogenic influences by several studies (e.g., [64, 65]). Superposed on the general warming are interannual and decadal signals that can be caused by natural climate variability, including the impact of volcanism and variable solar radiation [65–68]. The spread among all series is considerable, however, and this can be an issue particularly if one is interested in studying their year-to-year evolution. Differences in the series can also affect the quantitative determination of long-term trends. Differences are not confined to the period before the 1990s when coverage is sparse (see Fig. 2) and methodologies to fill in data gaps can be influential. Other factors, including quality control and bias correction procedures, can affect the estimates, even over most well-covered periods [41].

Recent discovery of several systematic biases in XBT and Argo data [13, 40, 69] has led to various updates in mean heat content series. The estimates in Fig. 5



Heat Content and Temperature of the Ocean. Figure 5
Anomalies in upper ocean (0–700 m) heat content relative to the mean over the period 1955–2002. Various curves are from several data-based estimates (Reproduced from Fig. 2 of Palmer et al. [63]). The spread among all the curves is caused by several factors, including the specific data used, the applied bias corrections, and the objective analysis methods for smoothing and interpolation



Heat Content and Temperature of the Ocean. Figure 6 Linear trends for heat content in the upper ocean (0–700 m) reproduced from Fig. 2 of Carton and Santorelli [52] and corresponding to several different analyses for approximately the period 1960–2005. Apart from the "Levitus" and "Ishii" fields, which are purely derived from observations, all others involve ocean models and data assimilation schemes. Details of all the analyses are given in Carton and Santorelli [52] and references therein. Coloring represents regions where values are outside the range ± 1.5 × 10^8 J/m^2/decade

include several of these bias corrections, which imply major differences in temporal evolution with previously published results (e.g., [70]). Efforts to understand the differences in Fig. 5 are ongoing (e.g., [41, 63]) and likely to continue in the foreseeable future. Multi-instrument approaches to estimating heat content, such as combining in situ temperature and altimeter sea level [29] or comparing XBT and CTD temperature estimates [69] can be useful when trying to rid the observing system of systematic errors in the various data streams, which are difficult to detect otherwise. Ultimately the synthesis of all data sets using models and data assimilation can bring out similar benefits (e.g., [52, 71]).

If knowledge of spatial mean temperature and heat content variability, with all the averaging and noise reduction involved, suffers from significant uncertainties, robust information on spatial patterns of variability is even more problematic because of very poor data coverage in extensive areas such as the Southern Ocean (Fig. 2). Even for the more recent Argo period, significant observational gaps are still present in ice-covered regions and marginal seas (Fig. 3). Some of the earliest estimates of variability were based on differences over a few pentadal periods or on time series for whole oceanic basins, to allow for meaningful averages [72, 73]. More recently, state estimation/data assimilation methods have been brought to bear on the problem (e.g., [48, 74]).

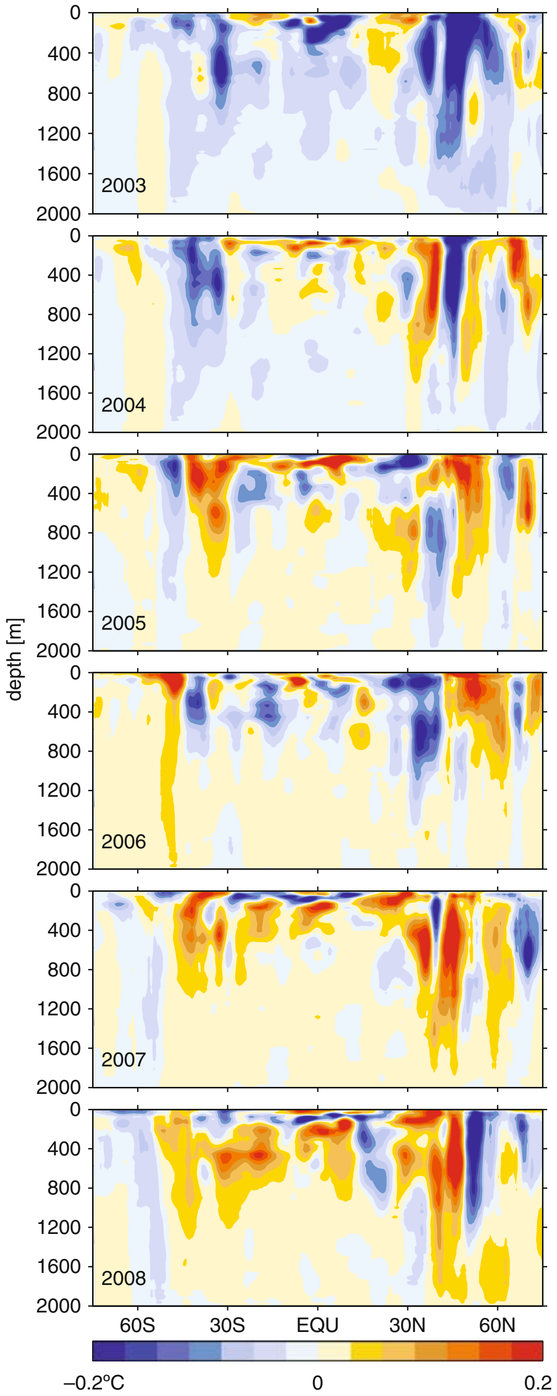
Long-term trends in upper ocean heat content (0–700 m), calculated over approximately the last three decades from a number of different analyses and methods summarized in Carton and Santorelli [52], are shown in Fig. 6. Although the warming trend in global mean heat content is clear in Fig. 5, the spatial patterns in Fig. 6 include both places of warming and cooling. Complex spatial and time variability patterns in ocean temperature fields are also evident in the acoustic measurements analyzed by Dushaw et al. [34]. The oceans are far from warming at the same rate everywhere. Despite considerable differences among all analyses, which can be taken as a crude measure of their uncertainty, there is a general tendency for warming in places like the Atlantic and parts of the south and north Indian Ocean, as well as cooling over the equatorial Pacific and parts of the North Pacific.

The inhomogeneous patterns of warming and cooling in Fig. 6 reflect the complexity of air-sea heat transfers and oceanic heat transports, both in the horizontal and in the vertical. Aside from the long-term (multi-decadal) scales represented in the trends in Fig. 6, heat content variability at all time scales permeates the available records (e.g., [52]). As with SST, some of the spatial patterns in upper ocean heat content can be related to primary modes of natural climate variability reviewed by Deser et al. [53]. Changes in major circulation systems such as the meridional overturning circulation in the North Atlantic [66] or the Antarctic Circumpolar Current in the Southern Ocean [75] can be involved. Achieving a full understanding of the mechanisms responsible for past upper ocean heat content variability is hampered, however, by the poor observational base. As societal focus shifts to the determination and understanding of regional climate variability and how it affects human life at a local level, the need to determine spatial patterns of variability in heat content will only grow in the future. Maintaining some of the recently implemented in situ and satellite observing capabilities is a minimum requirement to be able to monitor heat content variability both at the regional and global level.

As the focus of analysis descends to layers below the XBT maximum sampling depth of 700 m, the amount of historical observations diminishes considerably. Only the most recent period after the implementation of the Argo system, from 2003 onward, provides reasonable near-global coverage up to depths of around 2,000 m. The short record length precludes a comprehensive treatment of variability at all climate scales of interest, but the density and quality of the data provide for good estimates of spatial variability, both horizontally and vertically. Analyses of Argo data by von Schuckmann et al. [61] for the period 2003–2008 and comparisons with mean climatological temperature values confirm long-term warming tendencies, particularly in the North Atlantic, found by Levitus [72], Levitus et al. [76], and Antonov et al. [77] and show that temperature variability on interannual and decadal time scales can reach to 2,000 m and deeper, well below the XBT sampling range. The spatial and time structure of interannual temperature signals is quite complex, however.

Figure 7 reproduced from von Schuckmann et al. [61] shows an example of year-to-year zonally averaged temperature anomalies for the Atlantic Ocean, given as a function of depth and latitude. Typical amplitudes range between ±0.2°C, alternating sign on relatively long (∼1,000 km) meridional scales. Aside from the long-term warming already noted, other time scales are involved. Spatially, anomalies in Fig. 7 can show vertical coherence, with the same sign extending over most of the sampled water column, or considerable vertical structure including reversals of sign with depth. In addition, strongest anomalies are not always found at the surface, and sizable variability is apparent at all sampled depths, with deepest variability present at mid- and high latitudes. A variety of processes can be related to the temperature signals in Fig. 7, including changes in the position of subtropical and subpolar temperature fronts, strong air-sea interaction and onset of deep convection in the high latitudes, and changes in the strength of the horizontal circulation. von Schuckmann et al. [61] discuss some of the mechanisms behind the variability in Fig. 7 with references to other relevant works. In any case, only now is the observational system becoming dense enough to begin putting together a detailed temporal evolution of the three-dimensional temperature fields. A full understanding of these observations will likely come with the accumulation of good quality data and dedicated modeling and data synthesis efforts in the future (e.g., [49, 78]).

Below the Argo sampling maximum depth of 2,000 m, not even the most recent years are well sampled at all. Yet these very poorly observed layers represent approximately half of the volume of the global oceans. Most of the current knowledge about basin-scale temperature structures and variability in the deep/abyssal oceans comes from high-quality full-depth CTD casts obtained as part of global programs like WOCE [16] and earlier campaigns such as those during the International Geophysical Year in the late 1950s (e.g., [51]). Some attempts at including depths up to 3,000 m in calculations of heat content have been



Heat Content and Temperature of the Ocean. Figure 7 Mean annual temperature anomalies for the period 2003–2008, calculated from Argo data as described in detail by von Schuckmann et al. [61]. Values represent zonal averages for the Atlantic basin (Reproduced from Fig. 9 of von Schuckmann et al. [61])

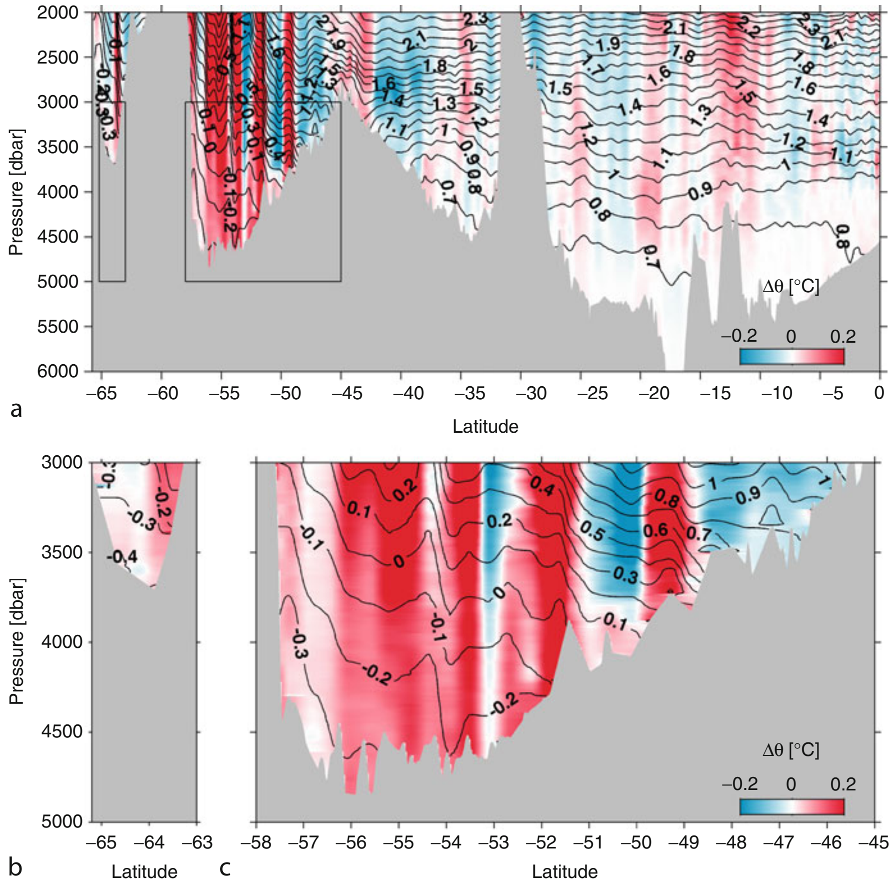
confined to highly averaged estimates in time and space (e.g., [76]). Most analyses, particularly for abyssal layers, tend to examine individual basin sections that have been reoccupied at various times, thus providing snapshots separated by several years to decades (e.g., [51, 79–82]). The observed changes in temperature are then used to infer variability, presumed to be long term and large scale in character, but great care must be taken to properly interpret such sparse observations.

Figure 8 shows an example of decadal temperature changes in the abyssal eastern South Indian Ocean from Johnson et al. [81]. Differences between temperature measurements taken in 1994–1995 and 2007 range between ±0.2°C, with strongest warming observed in latitudes south of 50S. Large variations are seen all the way to the bottom, well below 2,000 m. The warming in the Australian-Antarctic Basin below 3,500 m amounts to a mean change of +0.1°C and is statistically significant and generally consistent with other observational evidence [81]. Such a warming is equivalent to a heat flux imbalance of 0.9 W/m^2, which is substantially larger than values of 0.1 W/m^2 typically associated with geothermal heat flux at the ocean bottom [83]. Causes of the increase in temperature in Fig. 8 are thus likely related to changes in bottom circulation and water mass properties connected to surface processes around Antarctica, where bottom dense waters are formed through strong heat losses to the atmosphere and subsequently descend to great depths before spreading northward.

Abyssal temperature changes similar to those in Fig. 8 have been observed in many other ocean basins. For example, Fukusawa et al. [79], Kawano et al. [84], and Johnson et al. [80] document deep and abyssal warming across several sections of the North and South Pacific. Abyssal changes tend to be larger and decrease in range from south to north, with section mean changes in potential temperature as large as 0.01°C [80]. Similar warming of about 0.04°C is found in the bottom 1,500 m of the Brazil Basin in the South Atlantic [85]. As with the findings in the Indian Ocean shown in Fig. 8, these temperature changes in the abyssal Pacific and Atlantic oceans cannot be explained solely by effects of geothermal heat flux and are likely related to anomalies advected from the south, where sources of bottom water are located [29, 79, 80, 85].

Although changes in abyssal temperatures are relatively small compared to the changes in the upper ocean (cf. Figs. 7 and 8), these changes involve large volumes of water and can amount to substantial signals in global oceanic heat content. Estimates of multidecadal changes in global mean heat content for depths 0–3,000 m can differ by 25% from those based on the upper 700 m only (e.g., [76]). Contributions from abyssal depths may be smaller but not negligible. For example, if the observed abyssal warming in the western South Atlantic is reflective of global mean changes, it would imply decadal changes in heat content in the abyssal ocean close to 50% of those observed in the upper ocean [85]. Similarly, abyssal temperature changes and corresponding steric height effects could account for significant sea level signals on basin scales. Observed warming in the bottom waters of the South Indian Ocean implies changes in sea level as large as 4 cm over nearly a decade [81]. For comparison, global mean rates of sea level rise have been around 3 cm/decade in recent years (e.g., [86]).

Given that most deep and abyssal observations are confined to a few sections for each basin, the spatial extent of the observed long time scale signals at depths



Heat Content and Temperature of the Ocean. Figure 8
Mean values of potential temperature (*contours*) and differences in potential temperature between the periods 1994–1995 and 2007 as a function of latitude and pressure (in decibars, which are approximately equivalent to depth in meters). Section covers most of the Southern Hemisphere, from Antarctica to the equator, at longitudes between 82E and 95E to the west of Australia. Panels (**b**) and (**c**) zoom in on the *boxes* drawn in (**a**) corresponding to abyssal waters in the Princess Elizabeth Trough and the Australian-Antarctic Basin southern most region. Full details are provided in Johnson et al. [81] (Reproduced from Fig. 4 of Johnson et al. [81])

is not known. Thus, global means or even basin means are difficult to estimate [76, 82] and the impact of deep and abyssal temperature changes on large-scale heat content and sea level remains the subject of speculation. Besides the extremely sparse sampling in space, examining snapshots of the temperature field every few years or decades presents other sampling issues. Fluctuations over a range of time scales, from hours to decades, can be present even at depth. As can be seen in the example in Fig. 8, differences of any two temperature sections contain numerous relatively small-scale features, which are generally related to the presence of eddies with time scales of a few months or shorter and with temperature signatures reaching all the way to the bottom. This eddy "noise" is large enough to make much of the differences in Fig. 8 and in many other similar studies statistically insignificant (e.g., [80, 81]). Much denser observations in both space and time are needed to improve statistical robustness and be able to discern whether observed changes are truly long term or the effect of short time scales, which cannot be resolved in the present records. By the same token, for the deep ocean, dynamical interpolations and extrapolations of the data in time and space using model and estimation techniques assume increased importance [48, 49]. These efforts provide the ability to use information from all observations, not just temperature, including data collected at other depths, to provide "best" estimates of the temperature variability in space and time.

Future Directions

Significant investment at the international level will be needed to sustain the current array of in situ and satellite systems (primarily Argo and altimetry) relevant to ocean heat content and temperature monitoring. At the same time, the severe under-sampling of many regions of the global oceans, including the high latitudes covered by sea ice and the deep layers below 2,000 m not sampled by Argo system, continues to be a problem in the quest to achieve truly global measurements of ocean temperatures and heat content at appropriate temporal resolutions for climate monitoring and other purposes. As society grapples with the need to assess current climate conditions and predict their evolution on decadal and longer time scales, future developments in ocean monitoring capabilities gain increased significance. Several paths for action are foreseen, with some of the seeds of an improved observing and estimating system already in place.

The development of float technology promises to extend the capabilities of the Argo system to ice-covered and deep waters [18, 87]. At the same time, efforts are underway to improve sampling of the marginal and coastal seas (e.g., [88, 89]). In all these areas, apart from the use of autonomous Argo-type floats, other observing platforms and technologies are expected to contribute to the collection of temperature data. Programs that use marine mammals for profiling at some high latitude regions (e.g., [90]) could be expanded. Gliders [91, 92] can provide routine profiling with considerably more control on sampling patterns than Argo-type floats and technology improvements could lead to operational glider deployments in many regions in the future. Moored instrumentation at all depths will be involved, as part of calibration programs or monitoring of "choke" points where variability and transports of ocean properties such as heat and freshwater require specific in situ observations.

Besides temperature measurements, other observing systems indirectly relevant for the determination of ocean heat content will see substantial advances over the next decade. High-precision satellite altimetry has reached quasi-operational status with the launch of Jason-2 and the scheduled launch of Jason-3 for 2014. A number of other missions scheduled to fly over the next few years will deliver denser surface height observations in space and time [93]. Potential breakthrough sampling is the goal of the Surface Water and Ocean Topography mission, which is still in the initial stages of planning [94]. Similarly, measurements of the variable gravity field from space with higher accuracies than presently possible should complement developments in altimetry to deliver better estimates of full-depth integrated changes in heat content [31]. In terms of in situ observations, acoustic thermometry [32] has been successfully exploited in the North Pacific [34] and the Arctic [95]. Expanded acoustic measurements discussed by Dushaw et al. [33] promise to deliver much needed coverage of deep layers over basin scales. Such measurements can also contribute to the monitoring of temperature change and heat transport variability across important "choke points" in the ocean, as in the ongoing program in Fram Strait [33].

Within the coming years, concerted and stepped up efforts to create an integrated global observing system will deliver improved measurements of temperature as well as many other oceanic parameters (salinity, velocity, bottom pressure, etc.) which are all indirectly linked to the determination of ocean heat content. Appropriate synthesis of all these diverse measurements, providing interpolation and extrapolation of the observations under rigorous physical conservation principles, can only be achieved with the use of models and state estimation techniques (e.g., [78]). In the future, as is established practice with the atmosphere (e.g., [46]), global estimates of temperature and heat content will come from the implementation of such systems. Similar machinery can be applied to the efficient design of observing systems that maximize collection of relevant information for the purposes of climate monitoring, analysis, and prediction [96]. Glider technologies begin to provide the possibility of adaptive sampling procedures, under which instruments are continuously directed at making measurements at times and places deemed important for determining the evolution of a given parameter of interest.

The combination of developments in observing, modeling, and estimating systems promises to deliver a revolution in the way ocean temperatures and heat content are determined in the near future. The impact on many areas of ocean and climate science will be hard to miss, from making it possible to relate sea level change to increased heat transfer from the atmosphere or melting of ice sheets, to providing better understanding of the variable Earth's radiation balance and hydrological cycle.

Acknowledgments

The author is indebted to J. Carton, G. Johnson, J. Kennedy, J. Lyman, M. Palmer, K. von Schuckmann, and S. Walker for help with the figures reproduced here. The support of NASA and the National Oceanographic Partnership Program is gratefully acknowledged.

Bibliography

Primary Literature

1. Stommel H (1965) The Gulf stream: a physical and dynamical description, 2nd edn. University of California Press, Berkeley, 248 pp
2. Thomson CW (1877) The voyage of the Challenger, vol 2, The Atlantic. Macmillan, London, 396 pp
3. Wüst G, Defant A (1936) Atlas zur Schichtung und Zirkulation des Atlantischen Ozeans. Schnitte und Karten von Temperatur, Salzgehalt und Dichte. In: Wissenschaftliche Ergebnisseder Deutschen Atlantischen Expedition auf dem Forschungs-und Vermessungsschiff "Meteor" 1925–1927, 6: Atlas, 103 plates
4. Gill AE (1982) Atmosphere–ocean dynamics. Academic, New York, 662 pp
5. Sverdrup HU, Johnson MW, Fleming RH (1942) The oceans: their physics, chemistry and general biology. Prentice-Hall, New York
6. Talley LD (2007) In: Sparrow M, Chapman P, Gould J (eds) Hydrographic atlas of the world ocean circulation experiment (WOCE), vol 2, Pacific Ocean. International WOCE Project Office, Southampton
7. Baker DJ Jr (1981) Ocean instruments and experiment design. In: Warren BA, Wunsch C (eds) Evolution of physical oceanography. MIT Press, Cambridge
8. Spilhaus AF (1938) A bathythermograph. J Mar Res 1:95–100
9. Spilhaus AF (1940) A detailed study of the surface layers of the ocean in the neighborhood of the Gulf Stream with the aid of rapid measuring hydrographic instruments. J Mar Res 3:51–75
10. Snodgrass JM (1968) Instrumentation and communications. In: Brahtz JF (ed) Ocean engineering: goals, environment, technology. Wiley, New York, pp 393–477
11. Lyman JM, Johnson GC (2008) Estimating annual global upper-ocean heat content anomalies despite irregular in situ ocean sampling. J Climate 21:5629–5641
12. Hanawa K, Rual P, Bailey R, Sy A, Szabados M (1995) A new depth-time equation for Sippican or TSK T-7, T-6, and T-4 expendable bathythermographs (XBT). Deep-Sea Res I 42:1423–1451
13. Willis JK, Lyman JM, Johnson GC, Gilson J (2009) In situ data biases and recent ocean heat content variability. J Atmos Ocean Tech 26:846–852
14. Gregg MC (1998) Estimation and geography of diapycnal mixing in the stratified ocean. In: Imberger J (ed) Physical processes in lakes and oceans, vol 54. Amer Geophys Union, Washington, DC, pp 305–338
15. Thorpe SA (2004) Recent developments in the study of ocean turbulence. Ann Rev Earth Planet Sci 32:91–109
16. Siedler G, Church J, Gould J (2001) Ocean circulation and climate: observing and modelling the global ocean. Academic, Boston
17. Roemmich D, Johnson GC, Riser GCS, Davis R, Gilson J, Owens WB, Garzoli SL, Schmid C, Ignaszewski M (2009) The argo program: observing the global ocean with profiling floats. Oceanography 22:34–43
18. Freeland, H, Roemmich D, Garzoli S, LeTraon PY, Ravichandran M, Riser S, Thierry V, Wijffels S, Belbéoch M, Gould J, Grant F, Ignazewski M, King B, Klein B, Mork K, Owens B, Pouliquen S, Sterl A, Suga T, Suk M, Sutton P, Troisi A, Vélez-Belchi P, Xu J (2010) Argo – A decade of progress. In: Hall J, Harrison DE, Stammer D (eds) Proceedings of the OceanObs'09: sustained ocean observations and information for society, vol 2, Venice, 21–25 Sep 2009

19. Woodruff SD, Diaz HF, Kent EC, Reynolds RW, Worley SJ (2008) The evolving SST record from ICOADS. In: Brönnimann S, Luterbacher J, Ewen T, Diaz HF, Stolarski RS, Neu U (eds) Climate variability and extremes during the past 100 years, vol 33, Advances in global change research. Springer, Dordrecht, pp 65–83
20. Reynolds RW, Smith TM (1994) Improved sea surface temperature analyses using optimal interpolation. J Climate 7:929–948
21. Donlon C, Robinson IS, Casey KS, Vasquez-Cuervo J, Armstrong E, Arino O, Gentlemann CL, May DA, LeBorgne P, Piolle J, Barton I, Beggs H, Poulter DJS, Merchant CJ, Bingham A, Heinz S, Harris A, Wick GA, Emery B, Minnett PJ, Evans R, Llewellyn-Jones D, Mutlow CT, Reynolds R, Hawamura H, Rayner NA (2007) The global ocean data assimilation experiment high-resolution sea surface temperature pilot project. Bull Am Met Soc 88:1197–1213
22. McPhaden MJ, Busalacchi AJ, Cheney R, Donguy J-R, Gage KS, Halpern D, Ji M, Julian P, Meyers G, Mitchum GT, Niller PP, Picau J, Reynolds RW, Smith N, Takeuchi K (1998) The tropical ocean-global atmosphere (TOGA) observing system: a decade of progress. J Geophys Res 103:14169–14240
23. Bourlès B, Lumpkin R, McPhaden MJ, Hernandez F, Nobre P, Campos E, Yu L, Planton S, Busalacchi A, Moura AD, Servain J, Trotte J (2008) The PIRATA program: history, accomplishments, and future directions. Bull Am Met Soc 89:1111–1125
24. McPhaden MJ, Meyers G, Ando K, Masumoto Y, Murty VSN, Ravichandran M, Syamsudin F, Vialard J, Yu L, Yu W (2009) RAMA: the research moored array for African-Asian-Australian monsoon analysis and prediction. Bull Am Meteorol Soc 90:459–480
25. Send U, Weller RA, Wallace D, Chavez F, Lampitt RL, Dickey T, Honda M, Nittis K, Lukas R, McPhaden M, Feely R (2010) OceanSITES. In: Hall J, Harrison DE, Stammer D (eds) Proceedings of OceanObs'09: sustained ocean observations and information for society, vol 2, Venice, 21–25 Sep 2009. ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.78
26. Michaels AF, Knap AH (1996) Overview of the U.S. JGOFS BATS and hydrostation S program. Deep-Sea Res 43(2–3):157–198
27. Karl DM, Lukas R (1996) The Hawaii ocean time-series (*HOT*) program: background, rationale, and field implementation. Deep-Sea Res II 43:129–156
28. White WB, Tai C-K (1995) Inferring interannual changes in global upper ocean heat storage from TOPEX altimeter. J Geophys Res 100:24943–24954
29. Willis JK, Roemmich D, Cornuelle B (2004) Interannual variability in upper ocean heat content, temperature, and thermosteric expansion on global scales. J Geophys Res 109:C12036. doi:10.1029/2003JC002260
30. Ponte RM (1999) A preliminary model study of the large-scale seasonal cycle in bottom pressure over the global ocean. J Geophys Res 104:1289–1300
31. Jayne SR, Wahr JM, Bryan FO (2003) Observing ocean heat content using satellite gravity and altimetry. J Geophys Res 108:3031. doi:10.1029/2002JC001619
32. Munk WH, Forbes AMG (1989) Global ocean warming: an acoustic measure? J Phys Oceanogr 19:1765–1778
33. Dushaw B, Au W, Beszczynska-Möller A, Brainard R, Cornuelle B, Duda T, Dzieciuch M, Fahrbach E, Forbes A, Freitag L, Gascard J-C, Gavrilov A, Gould J, Howe B, Jayne S, Johannessen OM, Lynch J, Martin D, Menemenlis D, Mikhalevsky P, Miller JH, Munk WH, Nystuen J, Odom R, Orcutt J, Rossby T, Sagen H, Sandven S, Simmen J, Skarsoulis E, Stephen R, Vinogradov S, Wong KB, Worcester PF, Wunsch C (2010) A global ocean acoustic observing network. In: Hall J, Harrison DE, Stammer D (eds) Proceedings of OceanObs'09: sustained ocean observations and information for society, vol 2, Venice, 21–25 Sep 2009. ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.25
34. Dushaw BD, Worcester PF, Munk WH, Spindel RC, Mercer JA, Howe BM, Metzger K Jr, Birdsall TG, Andrew RK, Dzieciuch MA, Cornuelle BD, Menemenlis D (2009) A decade of acoustic thermometry in the North Pacific Ocean. J Geophys Res 114: C07021. doi:10.1029/2008JC0051
35. Boyer TP, Antonov JI, Baranova OK, Garcia HE, Johnson DR, Locarnini RA, Mishonov AV, O'Brien TD, Seidov D, Smolyar IV, Zweng MM (2009) World ocean database 2009. In: Levitus S (ed) NOAA atlas NESDIS 66. US Gov Printing Office, Washington, DC, p 219, DVDs
36. Johnson DR, Boyer TP, Garcia HE, Locarnini RA, Baranova OK, Zweng MM (2009) World ocean database 2009 documentation. In: Levitus S (ed) NODC internal report 20. US Government Printing Office, Washington, DC, p 175
37. Ishii M, Kimoto M (2009) Reevaluation of historical ocean heat content variations with time varying XBT and MBT depth bias corrections. J Oceanogr 65:287–299
38. Gouretski VV, Koltermann KP (2004) WOCE global hydrographic climatology [CD-ROM], Ber Bundesamt Seeschiffahrt Hydrogr Rep 35. Bundesamt Seeschiffahrt Hydrogr, Hamburg, pp 52
39. Worthington LV (1981) The water masses of the world ocean: some results of a fine-scale census. In: Warren BA, Wunsch C (eds) Evolution of physical oceanography. MIT Press, Cambridge
40. Gouretski VV, Kolterman KP (2007) How much is the ocean really warming? Geophys Res Lett 34:L01610. doi:10.1029/2006GL027834
41. Lyman JM, Good SA, Gouretski VV, Ishii M, Johnson GC, Palmer MD, Smith DM, Willis JK (2010) Robust warming of the global upper ocean. Nature 465:334–337
42. Bretherton FP, Davis RE, Fandry CB (1976) Technique for objective analysis and design of oceanographic experiments applied to MODE-73. Deep-Sea Res 23:559–582
43. Daley R (1991) Atmospheric data analysis. Cambridge University Press, Cambridge, 457 pp
44. Locarnini RA, Mishonov AV, Antonov JI, Boyer TP, Garcia HE (2010) World ocean atlas 2009, volume 1: Temperature. In: Levitus S (ed) NOAA Atlas NESDIS 68. U.S. Government Printing Office, Washington, DC, pp 184
45. Wunsch C (1996) The ocean circulation inverse problem. Cambridge University Press, Cambridge

46. Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteor Soc 77:437–470
47. Carton JA, Giese BS, Grodsky SA (2005) Sea level rise and the warming of the oceans in the simple ocean data assimilation (SODA) ocean reanalysis. J Geophys Res 110:C09006. doi:10.1029/2004JC002817
48. Wunsch C, Ponte RM, Heimbach P (2007) Decadal trends in sea level patterns: 1993–2004. J Climate 20:5889–5911
49. Köhl A, Stammer D, Cornuelle BD (2007) Interannual to decadal changes in the ECCO global synthesis. J Phys Oceanogr 37:313–337
50. Hansen J, Nazarenko L, Ruedy R, Sato M, Willis J, Del Genio A, Koch D, Lacis A, Lo K, Menon S, Novakov Y, Perlwitz J, Russell G, Schmidt GA, Tausnev N (2005) Earth's energy imbalance: confirmation and implications. Science 308:1431–1435
51. Roemmich D, Wunsch C (1984) Apparent changes in the climatic state of the deep North Atlantic Ocean. Science 307:447–450
52. Carton JA, Santorelli A (2008) Global decadal upper-ocean heat content as viewed in nine analyses. J Climate 21:6015–6035
53. Deser C, Alexander MA, Xie S-P, Phillips AS (2010) Sea surface temperature variability: patterns and mechanism. Annu Rev Mar Sci 2:115–143
54. Rayner NA, Brohan P, Parker DE, Folland CK, Kennedy JJ, Vanicek M, Ansell TJ, Tett SFB (2006) Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 dataset. J Climate 19:446–469
55. Smith TM, Reynolds RW, Peterson TC, Lawrimore J (2008) Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006). J Climate 21:2283–2296
56. Thompson DWJ, Kennedy JJ, Wallace JM, Jones PD (2008) A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. Nature 453:646–649
57. Rayner NA, Kaplan A, Kent EC, Reynolds RW, Brohan P, Casey KS, Kennedy, JJ, Woodruff SD, Smith TM, Donlon C, Breivik LA, Eastwood S, Ishii M, Brandon T (2010) Evaluating climate variability and change from modern and historical SST observations. In: Hall J, Harrison DE, Stammer D (eds) Proceedings of OceanObs'09: sustained ocean observations and information for society, vol 2, Venice, 21–25 Sep 2009. ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.71
58. Chen D, Cane MA, Kaplan A, Zebiak SE, Huang DJ (2004) Predictability of El Niño over the past 148 years. Nature 428:733–736
59. Compo GP, Whitaker JS, Sardeshmukh PD (2006) Feasibility of a 100 year reanalysis using only surface pressure data. Bull Am Met Soc 8:175–190
60. Antonov JI, Levitus S, Boyer TP (2004) Climatological annual cycle of ocean heat content. Geophys Res Lett 31:L04304. doi:10.1029/2003GL018851
61. von Schuckmann K, Galliard F, Le Traon P-Y (2009) Global hydrographic variability patterns during 2003–2008. J Geophys Res 114. doi:10.1029/2008JC005237
62. Levitus S, Antonov JI, Boyer TP, Locarnini RA, Garcia HE, Mishonov AV (2009) Global ocean heat content 1955–2008 in light of recently revealed instrumentation problems. Geophys Res Lett 36:L07608. doi:10.1029/2008GL037155
63. Palmer MD, Antonov J, Barker P, Bindoff N, Boyer T, Carson M, Domingues CM, Gille S, Gleckler P, Good S, Gourtetski V, Guinehut S, Haines K, Harrison DE, Ishii M, Johnson GC, Levitus S, Lozier MS, Lyman JM, Meijers A, von Schuckmann K, Smith D, Wijffels S, Willis J (2010) Future Observations for Monitoring Global Ocean Heat Content. In: Hall J, Harrison DE, Stammer D (eds) Proceedings of OceanObs'09: sustained ocean observations and information for society, vol 2, Venice, 21–25 Sep 2009. ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.68
64. Levitus S, Antonov JL, Wang J, Delworth TL, Dixon KW, Broccoli AJ (2001) Anthropogenic warming of earth's climate system. Science 292:267–270
65. Barnett TP, Pierce DW, AchutaRao KM, Gleckler PJ, Santer BD, Gregory JM, Washington WM (2005) Penetration of human-induced warming into the world's oceans. Science 309:284–287
66. Lozier MS, Leadbetter S, Williams RG, Roussenov V, Reed MSC, Moore NJ (2008) The spatial pattern and mechanisms of heat content change in the North Atlantic. Science 319:800–803
67. Church JA, White NJ, Arblaster JM (2005) Significant decadal-scale impact of volcanic eruptions on sea level and ocean heat content. Nature 438:74–77
68. White WB, Dettinger MD, Cayan DR (2003) Sources of global warming of the upper ocean on decadal period scales. J Geophys Res 108:3248. doi:10.1029/2002JC001396
69. Wijffels SE, Willis J, Domingues CM, Barker P, White NJ, Gronell A, Ridgway K, Church JA (2008) Changing eXpendable BathyThermograph fallrates and their impact on estimates of thermosteric sea level rise. J Climate 21:5657–5672. doi:10.1175/2008JCLI2290.1
70. Domingues CM, Church JA, White NJ, Gleckler PJ, Wijffels SE, Barker PM, Dunn JR (2008) Improved estimates of upper-ocean warming and multi-decadal sea-level rise. Nature 453:1090–1093. doi:10.1038/nature07080
71. Stammer D, Köhl A, Awaji T, Balmaseda M, Behringer D, Carton J, Ferry N, Fischer A, Fukumori I, Giese B, Haines K, Harrison DE, Heimbach P, Kamachi M, Keppenne C, Lee T, Masina S, Menemenlis D, Ponte R, Remy E, Rienecker M, Rosati A, Schröter J, Smith D, Weaver A, Wunsch C, Xue Y (2010) Multi-model ensemble ocean synthesis in support of climate diagnostics. In: Hall J, Harrison DE, Stammer D (eds) Proceedings of OceanObs'09: sustained ocean observations and information for society, vol 2, Venice, 21–25 Sep 2009. ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.85
72. Levitus S (1989) Interpentadal variability of temperature and salinity at intermediate depths of the North Atlantic Ocean, 1970–74 versus 1955–59. J Geophys Res 94:6091–6131

73. Levitus S, Antonov J, Boyer TP, Stephens C (2000) Warming of the world ocean. Science 287:2225–2229
74. Köhl A, Stammer D (2008) Decadal sea level changes in the 50-year GECCO ocean synthesis. J Climate 21:1876–1890
75. Böning CW, Dispert A, Visbeck M, Rintoul SR, Schwarzkopf FU (2008) The response of the Antarctic circumpolar current to recent climate change. Nature Geosci 1:864–869
76. Levitus S, Antonov JI, Boyer TP (2005) Warming of the world ocean, 1955–2003. Geophys Res Lett 32:L02604. doi:10.1029/2004GL021592
77. Antonov JI, Levitus S, Boyer TP (2005) Thermosteric sea level rise, 1955–2003. Geophys Res Lett 32:L12602. doi:10.1029/2005GL023112
78. Wunsch C, Heimbach P, Ponte RM, Fukumori I, The ECCO-GODAE Consortium Members (2009) The global general circulation of the ocean estimated by the ECCO-consortium. Oceanography 22:88–103
79. Fukasawa M, Freeland H, Perkin R, Watanabe T, Uchida H, Nishina A (2004) Bottom water warming in the North Pacific Ocean. Nature 427:825–827. doi:10.1038/nature02337
80. Johnson GC, Mecking S, Sloyan BM, Wijffels SE (2007) Recent bottom water warming in the Pacific Ocean. J Climate 20:5365–5375. doi:10.1175/2007JCLI1879.1
81. Johnson GC, Purkey SG, Bullister JL (2008) Warming and freshening in the abyssal southeastern Indian Ocean. J Climate 21:5351–5363. doi:10.1175/2008JCLI2384.1
82. Purkey SG, Johnson GC (2010) Warming of global abyssal and deep Southern Ocean waters between the 1990s and 2000s: contributions to global heat and sea level rise budgets. J Climate 23:6336–6351
83. Stein C, Stein S (1992) A model for the global variation in oceanic depth and heat flow with lithospheric age. Nature 359:123–129
84. Kawano T, Fukasawa M, Kouketsu S, Uchida H, Doi T, Kaneko I, Aoyama M, Schneider W (2006) Bottom water warming along the pathway of lower circumpolar deep water in the Pacific Ocean. Geophys Res Lett 33:L23613. doi:10.1029/2006GL027933
85. Johnson GC, Doney SC (2006) Recent western South Atlantic bottom water warming. Geophys Res Lett 33:L14614. doi:10.1029/2006GL026769
86. Cazenave A, Nerem RS (2004) Present-day sea level change: observations and causes. Rev Geophys 42:RG3001. doi:10.1029/2003RG000139
87. Klatt O, Boebel O, Fahrbach E (2007) A profiling float's sense of ice. J Atmos Ocean Technol 24(7):1301–1308
88. Poulain P, Barbanti R, Font J, Cruzado A, Millot C, Gertman I, Griffa A, Molcard A, Rupolo V, LeBras S, Petit de la Villeon L (2007) MedArgo: a drifting profiler program in the Mediterranean Sea. Ocean Sci 3:379–395
89. Malone T, DiGiacomo P, Muelbert J, Parslow J, Sweijd N, Yanagi T, Yap H, Blanke B (2010) Building a global system of systems for the coastal ocean. In: Hall J, Harrison DE, Stammer D (eds) Proceedings of OceanObs'09: sustained ocean observations and information for society, vol 2, Venice, 21–25 Sep 2009. ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.59
90. Charrassin JB, Hindell M, Rintoul SR, Foquet F, Sokolov S, Biuw M, Costa D, Boehme L, Lovell P, Colman R, Timmermann R, Meijers A, Meredith M, Park Y-H, Bailleul F, Goebel M, Tremblay Y, Bost C-A, McMahon CR, Field IC, Fedak MA, Guinet C (2008) Southern Ocean frontal structure and sea-ice formation rates revealed by elephant seals. Proc Nat Aca Sci 105:11634–11639. doi:10.1073/pnas/0800790105
91. Rudnick DL, Davis RE, Eriksen CC, Fratantoni DM, Perry MJ (2004) Underwater gliders for ocean research. J Mar Tech Soc 38:73–84
92. Eriksen CC, Perry MJ (2009) The nurturing of seagliders by the National Oceanographic Partnership Program. Oceanography 22:146–157
93. Wilson S, Parisot F, Escudier P, Fellous J-L, Benveniste J, Bonekamp H, Drinkwater M, Fu L, Jacobs G, Lin M, Lindstrom E, Miller L, Sharma R, Thouvenot E (2010) Ocean surface topography constellation: the next 15 years in satellite altimetry. In: Hall J, Harrison DE, Stammer D (eds) Proceedings of OceanObs'09: sustained ocean observations and information for society, vol 2, Venice, 21–25 Sep 2009. ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.92
94. Durand M, Fu L-L, Lettenmaier DP, Alsdorf DE, Rodríguez E, Esteban-Fernandez D (2010) The surface water and ocean topography mission: observing terrestrial surface water and oceanic submesoscale eddies. Proc IEEE 98(5):766–779
95. Gavrilov AN, Mikhalevsky PN (2006) Low frequency acoustic propagation loss in the Arctic Ocean: results of the arctic climate observations using underwater sound experiment. J Acoust Soc Am 119:3694–3706
96. Heimbach P, Forget G, Ponte RM, Wunsch C, Balmaseda M, Awaji T, Baehr J, Behringer D, Carton J, Ferry N, Fischer A, Fukumori I, Giese B, Haines K, Harrison DE, Hernandez F, Kamachi M, Keppenne C, Köhl A, Lee T, Menemenlis D, Oke P, Remy E, Rienecker M, Rosati A, Smith D, Speer K, Stammer D, Weaver A (2010) Observational requirements for global-scale ocean climate analysis: lessons from ocean state estimation. In: Hall J, Harrison DE, Stammer D (eds) Proceedings of OceanObs'09: sustained ocean observations and information for society, vol 2, Venice, 21–25 Sep 2009, ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.42

Books and Reviews

Hall J, Harrison DE, Stammer D (eds) (2010) Proceedings of OceanObs'09: sustained ocean observations and information for society, vol 2, Venice, 21–25 Sep 2009. ESA Publication WPP-306

Koblinsky C, Smith N (eds) (2001) Ocean observations for the 21st century. GODAE Office/BoM, Melbourne

Peixoto JP, Oort AH (1992) Physics of climate. American Institute of Physics, New York

Sparrow M, Chapman P, Gould J (2007) The world ocean circulation experiment (WOCE) hydrographic atlas, 4 volumes. International WOCE Project Office, Southampton

Warren BA, Wunsch C (1981) Evolution of physical oceanography. MIT Press, Cambridge

High Speed Rail, Technology Development of

Eda Beyazit, Moshe Givoni
Transport Studies Unit, School of Geography and the Environment, University of Oxford, Oxford, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Technological Development of HSTs
The Spatial Development of HST
Main Planning Principles of HSTs
Economic Costs of HSTs and Financing Strategies
Impacts of HSTs
Future Directions: Future of the High-Speed Train
Bibliography

Glossary

Dedicated lines In transport and telecommunication, dedicated lines are special rail, road infrastructure, cable or other facilities dedicated to a specific application. In transport technology examples can be found in bus, train, and subway systems running on dedicated lines.

Induced traffic/demand Induced traffic (or demand) is the volume of traffic that is drawn to a new road by additional capacity [8]. Induced demand occurs from a number of different sources such as new infrastructure, capacity increase, and infrastructure improvement or land use and location changes.

Shinkansen Shinkansen (new trunk line in Japanese) refers to the high-speed trains in Japan but also to the world's first HSR corridor, Tokaido Shinkansen, which was opened between Tokyo and Osaka in 1964. The main reason for the development of the Shinkansen network was to increase the network capacity between major cities of Japan. Shinkansen runs on dedicated railways designed only for high speed.

TEN Trans-European Network (TEN) Program comprises of a large number of projects in telecommunications, transport, and energy [13]. Trans-European Transport (TEN-T) Network is an important part of this program aiming to establish and develop the key links and interconnections needed to eliminate existing bottlenecks to mobility fill in missing sections and complete the main routes – especially their cross-border sections cross natural barriers and improve interoperability on major routes [33]. TEN-T includes projects such as road, rail, seaport, inland waterway, airport, shipping management, air traffic management and navigation networks. Number of total rail projects makes almost 50% of all transport projects and high-speed rail network is one of the most important projects of TEN-T. With projects such as the PBKAL (Paris-Brussels-Köln-Amsterdam-London) network and Southwest Europe network, high-speed rail lines are expected to create substantial reduction in travel times, increase the capacity and promote alternative transport to air and road transport.

TGV TGV (Train à Grande Vitesse) is the name given to the high-speed trains in France, which later on adopted by other countries using the same technology. French TGV is the first high-speed train in Europe, started to serve between Paris and Lille in 1981. Unlike Shinkansen, TGV operates on conventional tracks as well, thus, it can serve in city centers as well as regions by reducing its speed and this leads to a better connectivity between regional and urban network and significant cost savings. French recorded the highest speed ever reached on *wheel on steel* trains as high as 574 kph in 2007 during test drives [34].

Tunneling Effects A notion brought up by Graham and Marvin [18] referring to the possible effects of transport systems connecting specific areas in a region and bypassing others. This leads to some highly mobile socio-economic groups traveling through these places without even noticing the existence of them. Connecting places via high-speed systems may create adverse impacts as it may push disconnected places to the periphery (see also space-time compression by Harvey [21]).

Definition of the Subject

High-Speed Rail is an advanced railway technology for carrying passengers at high speeds. It can be defined as the infrastructure capability to support high-speed rolling stocks. In the transport literature, rail (or railway) and train are often used synonymously [15]. In this section of the Encyclopedia of Sustainability the term will be used as High-Speed Train with reference to the technological development of both the vehicle (the train) and the track (the railways) but also with a reference to the *train* as a carrier and high-speed train not as a single element but as a system. Moreover, the main focus is on the passenger transport, although some emphasis will be made on the freight transport. Late twentieth century witnessed the return of the train as an important passenger transport mode and the reason for this is the rapid development of the high-speed rail technology. Twenty-first century is witnessing even a larger impact of train as a transport mode. World's biggest economies are racing with each other to implement high-speed train systems in order to enhance their economies with fast transfer between major centers of business, finance, residential development, and tourism. However, these ambitions come with a price: high-speed train can enhance economies and societies but at the same time it can create environmental, spatial, and socio-economic deprivation. This review evaluates the challenges that the future HST development faces within the context of sustainable development.

Introduction

> ▸ Transport technologies seldom make a comeback, save in nostalgia trips for well-heeled tourists. Stagecoaches have not made a reappearance on the Bath Road, nor sedan chairs on the streets of London. But there is a spectacular exception: railways, written off thirty years ago as a Victorian anachronism destined to atrophy before the steady growth of motorway traffic, have suddenly become one of the basic technologies of the twenty-first century.
>
> The reason of course is the high-speed train...
>
> (Hall and Banister 1993 [19, p. 157])

Speed has been one of the biggest ambitions of the mankind, not only in transport but also in different aspects of life: fast broadband, fast food, fast vehicles, and fast monetary transactions are only a few examples for the desire of a fast life. The reason behind this passion can be explained sociologically and can be attributed to the way in which modernism was shaped around the Fordist ideas of production and daily life in the twentieth century, however, one can discuss that it dates back to pre-Fordism, to the industrial revolution when the train emerged as one of the most important modes of transport of the time. Although rail lost its power over the years, especially after the introduction of the motorcar, in the last 40 years the number of passengers using rail has increased substantially within the European Union (EU15 see [14]) and in Japan. Nevertheless, the situation differs from one country to another. According to the figures of European Commission's 2010 Statistical Pocketbook, the Eastern European countries witnessed a significant decline in the rail use over the years while in the Western European countries the rail use increased significantly within the same period [12]. This might be related to the differences between countries in terms of "investment in rail services, the quality and frequency of services offered, the growth in car ownership, levels of expendable income, and the location of rail services with respect to the population" [14, p. 4].

High-Speed Train development can also be seen as one of the reasons of the differences between rail uses in countries since majority of the Western European countries have HST services, whereas HST development has recently started in Eastern Europe. However, HSTs still do not accommodate the majority of the trips in many countries. In countries where the HST growth took place more effectively, such as France, Germany, Sweden, Spain, and Italy, in 2008 share of HST in total rail transport was 61.9%, 28.5%, 27.2%, 22.9%, and 17.8%, respectively [12]. However, within the European Union, HST network accounts only 23.9% of the EU27 passenger kilometer in 2006 [12]. Although these figures show that the HST use in Europe is only a small fraction of the total rail usage and is mainly applied by the Western European countries, more and more countries are becoming interested in HST systems. By 2025 the HST network around the world will reach more than 40,000 km and one third of the whole network will be in China, a new emerging force in HST development ([36], See Fig. 4, [4]). Moreover, the EU White Paper for 2010 emphasizes the importance of an effective passenger high-speed rail network integrated with the air travel [10]. Recently, in the USA, the government committed $8 billion to

HST projects on 13 corridors across 31 states. All these investments around the globe shows that the governments started to become more aware of the benefits of a HST network as it not only increases the passenger capacity between places but it also can be a driving force of development in many scales from local to regional and national.

However, this fast development of HST network creates new challenges, especially in terms of the goals of sustainability. As discussed earlier, speed is one of the main elements of defining the movement of people and goods and in transport research travel time savings has been one of the main concerns: "[t] ransport appraisal is almost totally (80%) dependent on the user benefits resulting from travel time savings and hence the overwhelming desire to speed up traffic" [2, p. 955]. Since traveling in high speed is very attractive not only for businesses but also for tourism and daily commuting, HST often increases the demand for traveling which is one of the main obstacles of sustainability and one of the main arguments of the sustainable mobility paradigm [1]. In transport literature, the volume of traffic created on a route with new infrastructure or expansion in the current road structure is called the *induced demand*. With the help of the new improvements, cars, passengers, and trucks are drawn from other routes or even new demand is generated with the relief in traffic on some routes. As Cervero [5] states "[…] road improvements, critics charge, provide only ephemeral relief—within a few years, most facilities are back to square 1, just as congested as they were prior to the investment" [5, p. 3]. This phenomenon points out one of the main challenges for HST development in terms of sustainability: "how should the railway take a larger share of the current demand for transport without inducing new demand" [14, p. 1].

Under the light of the discussions above, this review will first give explanation of the meaning of the High-Speed Train and evaluate the technological development of HSTs by demonstrating different HST types. It will then discuss the principles of HSTs briefly and mention the latest debates on the transport, socio-economic, and environmental impacts of HSTs. The review will be finalized with a discussion on the future of High-Speed Train system.

Technological Development of HSTs

The idea of increasing the speed of the railways dates back to as early as 1903 when in Germany test drives achieved 200 kph [9]. Since then and until the inauguration of the modern HST, tests continued in many countries. Minor technological attempts in Germany during 1930s were followed by the achievement of Italian ETR in 1938, reaching up to 203 kph during test drives. However, these tests were interrupted by the Second World War. The attempts took off again in 1955 in France where the highest speed by that time was recorded as 331 kph. Although the first attempts were made in Europe, it was Japan who first succeeded in the commercial operation of HSTs after its success in recording 256 kph in test drives in 1963.

Although much of the reference is to the increased speed while defining High-Speed Trains, the main reason for constructing such a system is to increase the capacity of routes. HST lines can increase the capacity first, by supplementing the existing lines with more capacity and second, by shortening the headway between trains. The latter generates as the frequency between services increases due to high speeds, longer trains and the most up-to-date signaling systems [15]. Higher frequency increases the capacity, therefore makes the HST an attractive investment for countries. For instance, Shinkansen carried 400,000 passengers between Tokyo and Osaka per day in 2010 [34].

It is difficult to talk about a single definition of High-Speed Train (HST), however, "the reference is always to passenger services and not to freight" [15, p. 594]. In order to define HST, one should look at the meaning of high speed: "[h]igh speed can relate to the infrastructure capability to support high speed, the rolling stock capability to achieve high speed and/ or the actual operation speed achieved" (ibid. 594). According to the European Union (EU), high speed is 250 kph for dedicated new lines and 200 kph for upgraded lines in respect of the infrastructure capabilities [38]. Therefore, EU describes High-Speed Train as;

- Those built specially for High-Speed travel.
- Those specially upgraded for High-Speed travel. They may include connecting lines, in particular junctions of new lines upgraded for High Speed with town center stations located on them, on which speeds must take account of local conditions.

Thus, high-speed lines would comprise:

- Specially built High-Speed lines equipped for speeds generally equal to or greater than 250 km/h.
- Specially upgraded High-Speed lines equipped for speeds of the order of 200 km/h.
- Specially upgraded High-Speed lines which have special features as a result of topographical, relief or town-planning constraints, on which the speed must be adapted to each case.

Currently, with some HSTs operating at speeds of 350 kph, 200 kph might not seem high speed anymore. "In practice, however, speed has not always been the best indicator, since commercial speed in many services is often limited due to, for example, proximity to densely urbanized areas or the existence of viaducts or tunnels" ([40], p. 20). Thus, speed cannot be regarded as the only determining factor of the HSTs and according to the International Union of Railways there are two important principles composed of different measurements [35]:

- High-Speed Railway is not element but systems. They are in fact very complex system, comprised by state of the art of the following components: Infrastructure, station emplacement, rolling stock, operation rules, signaling systems, marketing, maintenance systems, financing, management, legal aspects
- High speed is not unique and must be adapted to all countries and circumstances: As well as other railway systems, high speed can be considered under different conceptions, regarding commercial aspects, operation aspects, etc.

The differences between countries in terms of geographic and demographic characteristics, needs, and resources have led to the evolution of different types of HSTs [15]. These models which developed in the conditions of different countries then adopted by others, can be grouped into four: The Shinkansen, TGV, The tilting HST and The MAGLEV. However, in the last 10 years these models have been evolved to include each other's features. Examples such as tilting technology adopted by the TGV and Shinkansen trains can be found. Furthermore, MAGLEV technology is still in the first stages of its development and it is not certain whether this technology will be implemented widely in the future or will be completely abandoned due to extremely high construction costs compared to other HST models.

Shinkansen Model: The Bullet Train

Japanese Shinkansen is regarded as the first modern HST in operation [15]. The Tokaido Shinkansen opened in 1964 between Tokyo and Osaka as a celebration of the Olympics hosted by Japan [20]. However, the main reason was to increase the passenger capacity between these cities. Shinkansen travels 552.6 km, connecting Tokyo, Nagoya, and Osaka at a maximum speed of 270 kph in 2 h and 25 min. As of March 2010, approximately 138 million passengers are carried per annum [23] and there are 400,000 passengers on board per day [34]. Remarkably, since its opening no fatal accidents or injuries have been recorded [23].

Due to the unique geographic features of Japan and technical differences between conventional tracks and new gauges, it was not possible to achieve high speeds on the existing narrow gauge lines. Thus, the Shinkansen was built on a new line, dedicated only to high-speed trains. With this feature, Shinkansen differs from other HST models in most of Europe. The need for construction of new lines, high usage of tunnels and high land values within city centers are some of the factors that increase the building costs of Shinkansen [15] (Fig. 1).

High Speed Rail, Technology Development of. Figure 1 Shinkansen (Source: BakaOnigiri, Wikipedia)

The TGV

French TGV (Train à Grande Vitesse) In Europe, where the high-speed system emerged in its modern form in 1970–1980s, the rationality behind building high-speed lines and developing it in such a way that the lines in different countries would connect with each other and other transport networks, was to build up a Trans-European High-Speed Rail Network in the future. TEN-T network comprises of road, rail, air, and water transport networks designed to provide seamless transport within the European Union and High-Speed Rail network has a crucial role in achieving the cohesion goals of the EU. According to Vickerman, the reasons for creating such a network were [37, p. 22]:

- To overcome limited capacity on critical links of the rail network, where some new investment was needed and a completely new alignment appeared the most cost effective solution.
- To increase speeds on particularly slow sections of the trunk network where the developments in rail technology could not be fully exploited.
- To improve the accessibility of more remote regions.

However, it was initially France, within the EU members, where the high-speed lines were developed in a network and later on expanded to connect with other national lines.

Although French began test drives as early as 1950s, it was not until 1980s that they put the first modern HST into service in Europe. The first TGV line opened between Paris and Lyon in 1981. It serves the same purpose as Shinkansen but differs in design due to different features of France. "The most significant difference between the TGV and the Shinkansen is probably the ability of the former to operate on conventional tracks as well, which allows the TGV to use the conventional lines as it enters and leaves the city center, leading to significant cost savings" [15]. This feature of the TGV creates a network extending to regions without HST and allows connections between major cities and small towns where the demand is not high enough to justify the construction of a dedicated line (ibid.). This operational model is an example to the mixed high-speed model [40]. The TGV can reach up to 320 kph for commercial speed. French recorded the highest speed ever reached on *wheel on steel* trains as high as 574 kph in 2007 during test drives [34]. Nevertheless, operating at this speed is not possible at the moment (Fig. 2).

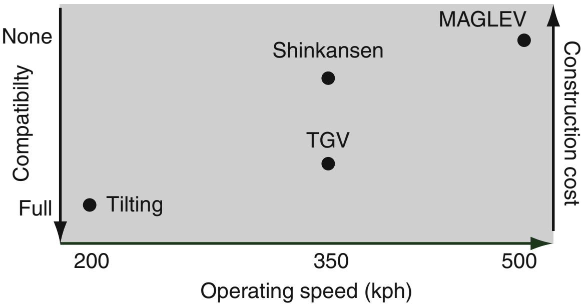
High Speed Rail, Technology Development of. Figure 2 TGV at Gare de Lyon (Source: Sese Ingolstadt, Wikipedia)

Since 1981, the opening of the first TGV line 1,400 million passengers have been carried [34]. Currently the French TGV consists of eight lines [35]: TGV Sud-Est (Paris – Lyon); TGV Rhône-Alpes (Lyon – Valence); TGV Méditerranée (Valence – Marseille); TGV Atlantiqué (Paris – Le Mans & Tours); TGV Interconnexion-Est (east of Paris); TGV-Nord (Paris – Lille –London via Channel Tunnel, Brussels – Amsterdam – Cologne); TGV Est (Paris – Strasbourg); TGV Peripignan Figures (France – Spain). The current network under operation is 1,896 km with a further 210 km under construction and 2,616 km planned (ibid.).

The Tilting Trains

Tilting train is a type of train with a sophisticated computerized mechanism running on conventional lines. The mechanism helps the train to tilt while it is going through the curves so that the train can travel on conventional lines at high speeds and passengers do not feel the discomfort from the centrifugal force. Tilting technology was initially invented in the UK in early 1980s, however, it was later abandoned. Nowadays, Italy is the leading country in tilting train technology with its Pendolino trains. Opposite to the Japanese, French, Spanish, and German HST trains which all use a newly built track on the sections where high speed is achieved (except the German ICE-T), in Italy the tilting technology is developed on routes where the demand is not high enough to justify the cost of constructing new tracks that allow high-speed operation [15].

The German ICE-T, the Swedish X-2000 and the Virgin Trains in the UK connecting London, Birmingham, Manchester, Liverpool, and Glasgow are the examples of tilting trains adopting the technology from Italy's Pendolino (ETR-450) trains. The Italian Pendolino reaches up to the maximum speed of 250 kph, whereas the Swedish X-2000 and the UK Virgin Trains Pendolino are slightly slower reaching up to 210 kph and 225 kph, respectively. Tilting mechanism is adopted by other HST models such as Shinkansen trains which use this technology to increase their speed.



High Speed Rail, Technology Development of. Figure 3 Operating speed, construction cost, and compatibility (with the conventional network) characteristics of the four high-speed train models ([15, p. 509], updated by the authors)

The MAGLEV

The idea of running the train on magnetic plates dates back to as early as 1970s, to the first tests conducted in Germany. "The technology relies on electromagnetic forces to cause the vehicle to hover above the track and move forward at theoretically unlimited speeds" [15, p. 597]. In practice, the aim is for an operation speed of 500 kph (Taniguchi [39] cited in [15, p. 597]. Japan is one of the leading countries in testing the MAGLEV technology. In 1997, tests started on the Yamanashi Maglev Test Line [24]. According to Japanese Railway Technical Research Institute that is responsible of conducting the MAGLEV tests in cooperation with the JR Central, in 2003, a MAGLEV three-car train set attained a maximum speed of 581 kph in a manned vehicle run [29]. In November 2004, two train sets succeeded in passing each other at relative speeds of up to 1,026 kph [29]. Even though the MAGLEV technology has been proved to be successful in the test drives, it has not been put into practice widely due to high construction costs and no compatibility with the existing lines. In Japan, Yamanashi Maglev Test Line has recently been agreed to be extended for another 23 km and is planned to become a part of the Tokaido Shinkansen Bypass, making it the first long-distance MAGLEV line in commercial operation [25]. In China, a short MAGLEV line was opened in December 2003 connecting Shanghai Airport and the city's Pudong financial district with trains running at maximum speed of 430 kph. However, as mentioned earlier, the investment on MAGLEV trains is very small in number compared to the rest of the HST development all around the world.

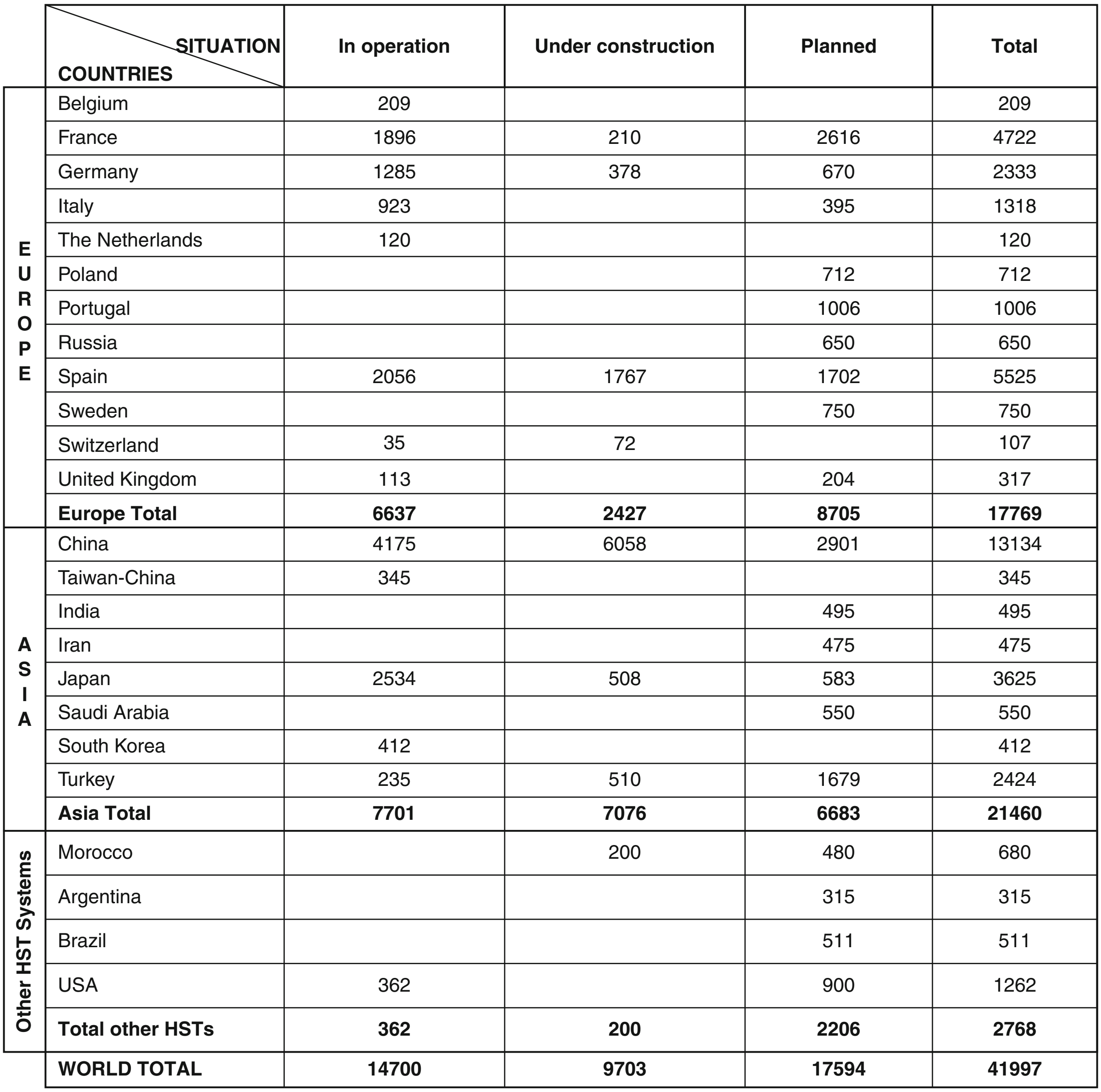
As explained above, there are differences in principle between the models of HSTs. Givoni [15] compared these models in terms of compatibility, construction costs, and speed (Fig. 3).

The Spatial Development of HST

According to the reports of the International Union of Railways [34], the HST lines in operation and construction cover many of the developed and developing countries. Nowadays there are high-speed train services in 14 countries and planned lines in other ten countries [36]. According to the UIC 2011 report, there is 14,700 km of high-speed line in operation, 9,703 km under construction and 17,594 km planned HST system in the world, that will make 41,997 km of high-speed line in 2025, 17,769 km of which in Europe and 21,460 km in Asia [36] (Fig. 4). However, considering more than 150 countries in the world have railway systems [28]; the number of countries investing in HST system is very low. Thus, it is questionable how much the HSTs can be a part of the solution for the increasing travel demand and whether it is the high costs or lack of planning or else that is making some countries reluctant of investing in this technology.

Europe

Besides France and Italy, HST development takes places in other European countries as well. The Spanish AVE



	SITUATION / COUNTRIES	In operation	Under construction	Planned	Total
EUROPE	Belgium	209			209
	France	1896	210	2616	4722
	Germany	1285	378	670	2333
	Italy	923		395	1318
	The Netherlands	120			120
	Poland			712	712
	Portugal			1006	1006
	Russia			650	650
	Spain	2056	1767	1702	5525
	Sweden			750	750
	Switzerland	35	72		107
	United Kingdom	113		204	317
	Europe Total	**6637**	**2427**	**8705**	**17769**
ASIA	China	4175	6058	2901	13134
	Taiwan-China	345			345
	India			495	495
	Iran			475	475
	Japan	2534	508	583	3625
	Saudi Arabia			550	550
	South Korea	412			412
	Turkey	235	510	1679	2424
	Asia Total	**7701**	**7076**	**6683**	**21460**
Other HST Systems	Morocco		200	480	680
	Argentina			315	315
	Brazil			511	511
	USA	362		900	1262
	Total other HSTs	**362**	**200**	**2206**	**2768**
	WORLD TOTAL	**14700**	**9703**	**17594**	**41997**

High Speed Rail, Technology Development of. Figure 4
The HST lines in operation, under construction and in the planning stage. Based on [36]

(Alta Velocidad Espanola, or Spanish High Speed), opened in 1992, can be seen as another type of TGV as it uses TGV-type rolling stocks in order to create integration with rest of Europe, mainly with France. However, like Shinkansen, "it runs on a dedicated line throughout, because the Spanish conventional network is wider than the standard International Union of Railways (UIC) gauge used across most of Europe" [15, p. 596]. Since it runs on dedicated lines and not compatible with the rest of the domestic network, it is possible to classify AVE in the Shinkansen category. In order to solve the problem of integration within the country, tracks have been developed to have three rails, thus, both types of trains (conventional and AVE) can run on them. The Spanish AVE achieved 350 kph in test drives and runs on 300 kph max speed commercially, connecting Madrid-Barcelona (689 km) in 2 h 38 m and Madrid – Sevilla (471 km) in 2 h 30 m.

Germany's HST, the ICE (Inter-City Express) that started to operate in 1991, today reaches the speeds of up to 320 kph [7]. Unlike other systems, the initial network was planned for both passenger and freight transport, using a technology deviating from the TGV and Shinkansen models. However, this mixed-use turned out to be a disadvantage due to high construction costs and low utilization of the lines, thus, no other country attempted to replicate it. Today the ICE network consists of different technologies with trains running on both dedicated lines and existing lines using the tilting technology where the demand is high enough to run high-speed train services but not high enough to construct dedicated lines [7]. Five different lines operate between several major European cities including international destinations in Switzerland, Austria, Belgium, and the Netherlands.

Portugal is also one of the leading countries in terms of the extent of the network to be developed. According to UIC [36], 1,006 km of new HST line has been planned to connect major cities with each other and with Spain. Portugal is followed by Sweden, Poland, and Russia that will have, respectively, 750 km, 712 km, and 650 km new high-speed lines in the near future. By the extension of 72 km HST lines in Switzerland and 204 in the UK, the HST lines in Europe (currently 6,637 km) will reach 17,769 km in 2025 [36]. In the UK, currently the only operating dedicated high-speed line is the Channel Tunnel Rail Link (CTRL, nowadays referred to HS1) which connects London to the Channel Tunnel.

Asia (Including Middle-East)

In Asia, following Shinkansen's success, China is aiming to be the new leader in HST development. Currently, there are 4,175 km of high-speed lines in operation which are planned to be extended to 13,161 km by 2025 [36]. When finished, China will have one third of the world's total HST network. Although South Korea (KTX – Korea Train eXpress – 412 km) and Taiwan (345 km) have put the HST lines into operation before China, they have no future plans to develop the network within the next 15 years. With the new lines in India, Iran, Saudi Arabia, and Turkey, the existing 7,701 km HST network in Asia will reach to 21,460 km in total in 2025 [36].

Other High-Speed Systems

In the USA, the only HST line in operation is the Acela Express tilting train running on the North East Corridor line between Boston and Washington, DC. Recently, the USA government announced the eight billion US dollars of state funds to be spent for the development of future HST lines. Currently, only the State of California is planning to develop a HST network between Los Angeles, San Francisco, and Sacramento which will add 900 km of HST line to the existing network [36]. In South America, Brazil, and Argentina planned 511 km and 315 km HST lines, respectively, which are aimed to be completed by 2020. Moreover, Morocco has 200 km of HST lines under construction and a further 480 km planned network [36].

Main Planning Principles of HSTs

As mentioned earlier, the main reason behind developing the high-speed rail technology is to increase the capacity of routes and reduce the travel time. However, increased speed is not solely sufficient to reach these goals; other measures need to be met as well. These are the integration of the HST network with the existing transport system, creating seamless transfer between the different modes of transport system and choosing the location and number of the stations.

First of all, there are three main models of station location: at the city center and in the outskirts (a third model, station at the edge of the city, can be included in either of the two models mentioned. For a detailed analysis and explanation of the subject, see Menéndez et al. [27]). Each option has its own pros and cons, and the balance is always context specific. From an accessibility perspective, the city center station location model might be better than the outskirts model as demand is likely to be higher due to higher population density and/or level of commercial activity. It is also likely to be much better served by public transport including regional and national train services. On the other hand, this model is significantly more costly and more complicated than the outskirts model due to restricted land for development and high costs of land intake. The outskirts model is likely to be cheaper in terms of land acquisition, simpler to construct and might offer more land for redevelopment and likely better connection opportunities for regional accessibility from a road transport perspective. Therefore, pros of one model become generally the cons of the other. Moreover, as Givoni and Banister mention the design of the stations is as important as the location in terms of allowing "seamless transfer between HST services and other rail (or air) services" [14, p. 9]. The transfer centers need to be designed in order to ease the passenger flows between different modes and also to accommodate different needs of passengers especially in international stations.

The number of stations is another important element in the planning of HSTs. In simple economic terms, the number of stations is a function of a trade-off between reducing the cost of accessing the station by opening an additional station, thus shortening the access journey, and the cost of adding a station which is the cost of providing it (construction) and the cost penalty of stopping the train (longer travel time) [17]. The characteristics of HST suggests it should have very low number of stops between the start and end stations. HSTs are feasible mainly on high-demand routes with stations located in major conurbations with 12–15 million passengers a year [37]. Most cities, except large metropolises like Tokyo and Paris, have one HST station. Increased number of stations means longer travel time, as the train will stop frequently and that would in turn reduce the desirability of the high-speed train. For instance, in France, an important number of trains between Paris and Lyon do not stop for 447 km which is different than the UK Channel Tunnel rail link that has three stops between London and the Channel tunnel which is 109 km apart ([20, 22]). However, there are also negative consequences of limiting the number of stations and planning them only for the major cities. "Since the HST often replaces or supplements an existing conventional railway service, these bypassed cities along the HST line will see substantial reduction in rail services because they will lose the through traffic to the HST and local demand would not support the same level of service. In turn, this reduction in rail accessibility can contribute to economic decline" [14, p. 6]. Thus, in routes where the demand is high enough to run extra services, cautions are taken due to possible negative impacts of HSTs. The different service levels can be found in Japan (see the "Socio-economic Impacts and Spatial Imbalance" section). However, as stated earlier, these measures can be applied in high-demand routes, therefore, for other places with low demand the integration of the HST network with the existing road, rail, and air system is essential.

There are three levels of integration: integration with the urban transport network, the national intercity transport network and the international, long-distance air transport network. The attractiveness of using HST compared to aircraft depends on the ease to get to the station vs. the airport and the time of doing so. Moreover, integration with the regional network is one of the essentials of HST development. Since the spatial spread of HST network is constrained by the station location choice and number of stations, its integration with the regional rail network is crucial for sustaining an increased accessibility from different regions. As Givoni and Banister suggest "the location of the HST stations must consider the entire rail network and not only local demand from cities" in order to spread the accessibility benefits from HST services across the entire network [14, p. 9]. Also, the integration with the rail network is more important than the road network as the latter would not create opportunities for a decrease in the car use due to the efficiency of the car in many ways such as door-to-door travel. Lastly, integration with air travel is another important strategy to spread the benefits of HST network. More emphasis will be made on the integration of air and rail in the next section.

It is those guiding principles for the development of HST that will determine the scope and scale of its impacts. Therefore, they are crucial in the planning of HST systems.

Economic Costs of HSTs and Financing Strategies

Construction costs and financing strategies of HSTs vary between countries and pretty much depend on the type of technology that is used. Different technologies are adopted by countries in order to overcome the restrictions that occur from geographic features or the existing technology of the railways.

The construction costs have been briefly discussed earlier when describing different HST models and different systems were compared in Fig. 3. However, a more detailed comparison of construction costs requires more information as the costs depend on so many criteria that are mentioned in the previous paragraph. Nevertheless, "the comparison of construction costs between different HSR projects is difficult since the technical solutions adopted in each case to implement these features do not only differ widely (depending on topography and geography), but also evolve over time" ([40], p. 22).

The UIC [41] points out three major types of costs for building a new HST structure. These are planning and land costs, infrastructure building costs, and superstructure costs [41]. According to a study comparing the construction costs of HSTs from different countries by Campos and de Rus [40], the range of costs varies between €9 and €39 million per kilometer (2005 figures). In Europe "France and Spain have slightly lower building costs than Germany, Italy, and Belgium. The HSR construction in Asia, except China, seems to be more expansive than in Europe" (ibid. 23). In addition to the building costs, operation costs are important in planning HST systems. According to Campos and de Rus [40], these costs can be grouped in two: infrastructure operating costs and rolling stock and train operating costs (for a detailed explanation of the operating costs see [40]).

Some of the well-known financing strategies can be discussed here. The TGV was financed by SNCF (Société Nationale des Chemins de fer Français – *French National Railway Corporation)* with expectations of minimum 12% financial rate of return. Unexpectedly, SNCF's investment was fully amortized by the end of 1993 after 12 years of service, proving that it was hugely profitable [37]. In 2010, SNCF's revenue from high-speed train services including TGV was €7.2 billion [30]. The SNCF is considered a state-owned industrial and commercial enterprise under French law [31]. Thus, the investment and operation of the TGV is under the control of the state. In Japan where the Shinkansen was initially constructed using government and World Bank funds the situation is slightly different. The Shinkansen lines were privatized in 1987 making the financing strategy a public-private partnership. The JNR (Japan National Railways) was divided to six passenger and one freight rail companies collectively called Japanese Railways Group (JR Group) that include East Japan Railway Company, West Japan Railway Company, and Central Japan Railway Company, all operating and maintaining different parts of the HST network. The initial investment for the Shinkansen was recovered by 1971, in 7 years of operation. Revenue from the Shinkansen has been an important source for financing local lines. However, these two examples (French TGV ad Shinkansen) are exceptional and the outcomes cannot be attributed to other HST systems.

Recently, in the US, the President Obama announced a spending of nearly $8billion for a new high-speed railway project funded by the Recovery Act in order to move the US in the direction of developing a high-speed rail service in 13 travel corridors covering 31 states all across this country with the goal of making high-speed rail accessible to 80% of Americans within 25 years [6]. In this sense, the financing strategy in the USA is similar to that of in France where the state controls the investment and operations through state-owned companies.

In the UK, the first high-speed line, CTRL, was initially planned as a state investment. However, due to the financial difficulties the governments signed several consortiums until 2010 when a 30-year concession to a consortium of Canadian investors was announced. For the HS2 the government set up an initiative under the name of HS2 Limited Company in 2009 for the design and development of the new HST line.

Impacts of HSTs

Transport Impacts

The main benefit of high-speed train which also fosters other benefits at economic and social levels is the increased route capacity. Here, the attribute is given not to the increased speed but to the reduced headways due to the high speeds. Higher speeds and the up-to-date signaling systems which are prerequisites of high-speed systems create short headway between trains and thus, trains run at higher frequencies [15] and provide more capacity on the routes where the existing network cannot meet the demand. Thus, changes in modal share occur due to the demand shift between transport modes and generation of additional demand. In TGV line between Paris and Lyon the share of the train journeys increased by 32% within the 3 years of its opening opposite to the decreased aircraft, car, and bus traffic (ibid.) “Paris-Lyon air traffic halved between 1980 and 1984, Paris-Geneve fell by around 20% and Paris-Sud Est in general grew at less than half the rate of other radial routes from Paris” [37, p. 24]. A similar result was seen in Spain after the inauguration of the AVE between Madrid and Seville where the train journeys increased by 35% between 1991 and 1994 [15]. Likewise, in the first 5 years of operation ICE, passengers more than doubled from just over 10 million to nearly 23 million [37]. “The AVE has been remarkably successful in terms of traffic growth and diversion. [. . .] [A]ir traffic has suffered with an estimated 60% reduction. Some 32% of travelers have diverted from air, 25% from car, and 26% were newly generated” (ibid. 30). However, it is important to mention that developing HSTs come at the expense of other railway networks. For instance, in EU 27 the number of passengers traveled by conventional railways showed no significant increase, whereas this increase is notable in HST trips (Fig. 5). This shows the fast increase in HST use while the rail use stays constant.

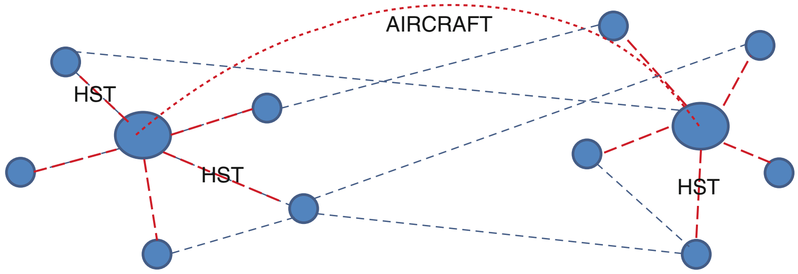
Apart from modal shift, high-speed lines also create substantial reduction in travel times which makes HSTs an attractive mode of transport. Before the inauguration of the HST in Japan, it took 7 h to travel between Tokyo and Osaka on the conventional line; it was then reduced to 4 h following the inauguration of the Shinkansen (Matsuda, 1993 cited by [15]). The current travel time between Tokyo and Osaka is 2 h 25 min with a maximum speed of 270 kph. Traveling with TGV between Paris and Lyon takes about 2 h at the moment. Likewise, in central and southern China almost 1,000 km is taken in 3 h and in northern China around 120 km is traveled in 30 min by Chinese HST, bringing countries’ largest cities closer to each other.

With reductions in travel time and changes in modal share, HSTs become competitors to airline companies. However, rather than competition, cooperation is more beneficial for airlines. “Cooperation between air and HST can take place if the airport is situated on the rail and HST networks allowing airlines to retain connecting traffic from short-haul routes to their more profitable long-haul services, thus giving up flights without compromising their Hub and Spoke

EU-27	HST		RAIL	
YEARS	Thousand million pkm	CHANGE	Thousand million pkm	CHANGE
1990	15,23		400,7	
1995	32,94	116%	350,5	–13%
2000	58,8	79%	370,7	6%
2005	80,11	36%	377	2%
2008	97,6	22%	409,2	9%

High Speed Rail, Technology Development of. Figure 5

Number of passengers transported by HST and conventional railways per km in EU27. Based on [12]



High Speed Rail, Technology Development of. Figure 6

Integration of HST network with airport can reduce short-haul flights operation" [14, pp. 6–7]. In Europe, Amsterdam Schiphol, Paris Charles de Gaulle, and Frankfurt International are examples where the seamless connection between the airport and high-speed train already exists.

Figure 6 demonstrates the possibility of airline-HST cooperation and integration of the HST network with the existing regional railway network. According to this, rather than providing services between minor cities, airlines can reduce the number of short-haul flights and operate on long-distance routes which are more profitable. The connection of airports with HST network which is also connected to the conventional rail network would not reduce the profits of the airline companies; on the contrary this type of connection would decrease their costs occurring from short-haul flights by keeping the number of long-haul flight passengers steady.

Sustainability Impacts

Socio-economic Impacts and Spatial Imbalance Transport investments that supply fast, affordable, reliable, and secure systems are seen as a condition for economic development either creating new economic areas with the help of agglomeration economies or increasing the size of actual markets (for a detailed understanding of the subject see [3]). "Development opportunities from new networks will occur at locations with good access to the new network and which provide nodal connections to other networks"[37, p. 35].

It is possible to observe this argument in the case of high-speed rail investments as well. Evaluating the previous studies on Shinkansen HST, Banister and Berechman [3] states that there are three conditions needed for growth: high incidence of information exchange industries; sufficient opportunities for higher education; good accessibility to a Shinkansen station [3, p. 279]. At the local level, Shinkansen's impacts occurred differently. The stations located in existing cities barely had any impacts, conversely, new stations at the edge of existing cities boosted the population and employment growth rates. First TGV line between Paris and Lyon created similar impacts in some parts on the route. A new station in Lyon Part-Dieu, close to central Lyon 2 h from Paris by TGV Sud-Est, has been largely affected from the opening of the TGV line especially in terms of the significant increase in the demand for office space (ibid.). Moreover, further local impacts have been observed as the TGV has been extended to the west and north. "Substantial growth has taken place at Le Mans, Nantes, and Vendôme in the TGV Atlantique" (ibid. 281). Also on the TGV-Nord line opened later in 1993 in order to connect Paris, London, and Brussels (northern Europe), a technopole concept, Euralille, was planned (ibid.). Initially, the network was planned strategically that Lille would become "a business bridge-head on the continent for London; city of services for Brussels and an attraction of relocations from Paris" ([42], p. 130). Recently, a major office development has been extended southward in Lille [20]. Likewise in Spain, AVE is known to have created new markets with the help of the shuttle services Ciudad Real and Puertollano around 200 km from Madrid [37].

However, new HST lines do not generate the same economic effects everywhere: "only those centers which are prepared to support it with complementary investment will stand to gain" [37, p. 35]. As seen in the example of The Channel Tunnel, where the stations are not located in major conurbations (but in the Kent countryside), the development potential accrues in other centers (ibid.), having maybe less impact that it would have had they been located in major centers. Nevertheless, a major office development around station, King's Cross and St Pancras after the completion of the CTRL (Channel Tunnel Rail Link) in 2007, can be seen [20].

The accumulation of investments only in major centers might not always have positive influences, especially when the subcenters or minor cities are concerned. As mentioned earlier, HSTs are feasible mainly on high-demand routes. This economic necessity makes it difficult for other areas – minor cities – to connect to the high-speed system which in turn creates economic and social imbalances. According to Graham and Marvin [18], "advanced infrastructure networks targeted on valued parts of the metropolis and drawing them into intense interaction with each other" create "tunnel effects" [18, p. 201]; also mentioned by Plassard, 1991 cited in [37]). In this case, HSTs connecting major cities within and beyond countries can create so-called tunneling effects, changing the proximity of places: making some closer to each other while pushing others to the periphery. Hall [20] calls this effect "peripheralization of the periphery" emphasizing that the regions which are already in the periphery, due to the lack of economic viability, are even more peripheralized as a result of making investments in other parts of the region. Maps of Europe as a "Shrinking Continent" have been created by Spiekermann and Wegener [32]. According to the time-space maps of rail travel times which are based on a deeper geographical understanding of space and time (see for example, [21] for more detailed understanding of the space-time compression), the core cities are drawn closer in terms of time-distance, while the more distant places which lack high-speed networks are bypassed and thus, become relatively more remote (ibid.). For instance, in a recent study by Martínez Sánchez-Mateos and Givoni [26] where the authors explore possible impacts of London – Birmingham HS2 link, they demonstrate that the towns in Birmingham catchment area will be affected from the construction of the HST stations in different ways. According to the study that explores the possible accessibility changes in the area after the introduction of the HS2 link, the towns that are relatively close to Birmingham and have good access to London will not benefit from the HS2, on the contrary, they might be adversely affected resulting in a decrease in their accessibility to London. Whereas, the settlements on the North of Birmingham that have less accessibility to London will benefit from the HS2 service more than the others. Thus, this would result in changing the geography of the county [26]. This example not only demonstrates the power of transport systems on geography but also shows the importance of station locations and integration between the transport modes.

In order to solve this problem, "in Japan the Shinkansen was deliberately structured on the basis of two service levels, fast (Hikari) serving only main cities and semi-fast (Kodama) serving intermediate cities, and this has been maintained; a super-fast (Nozomi) service was added in 1992" [20, p. 65]. In France and Germany similar measures were taken and additional stations were planned in small cities (ibid.). Likewise, in the UK, the Channel Tunnel Rail Link (HS1) was seen as a tool to regenerate the economy and Ebbsfleet and Stratford stations formed part of the Thames Gateway regeneration project east of London (ibid.). However, these solutions are only possible where the demand is high enough to compensate this investment. Furthermore, since high-speed trains are relatively more expensive compared to conventional trains, network solutions which aim to spread the impacts of HSTs would not be sufficient to make the system socially inclusive as not all socio-economic groups would be able to afford them. Here the crucial question is who benefits from such an expensive system.

Apart from economic and social impacts of HSTs that have been argued above, the safety benefits of HSTs are also important for the society. According to Campos and de Rus [40], "any comparison of accident statistics for the different transport modes immediately confirms that HSR is – together with air transport – the safest mode in terms of passenger fatalities per billion passenger-kilometers" (ibid. 25). As mentioned earlier, no fatal accidents or injuries on Shinkansen lines have been recorded since its opening [23].

Despite the social and economic inequalities that the HSTs might create in some regions, they are still important tools for the social and economic sustainability of the regions as they play an important role in the regeneration of the regions in the longer term and creating safe travel for societies. However, the governments need to be careful in planning these networks as like other transport investments they can be "the maker of some cities, but the breaker of the others" (Clark, 1957 cited in [20]).

Environmental Sustainability The environmental impacts of HSTs are not negligible. Although high-speed train systems are widely perceived to be climate friendly, there are serious environmental costs that they generate both during the construction and the operation. The impacts range from land take, barrier effects, visual intrusion, noise, and air pollution to contribution to global warming ([40]; [15]). The amount of energy used during construction and operation of HSTs, the source of energy and the level emissions created by the HSTs are important determinants of how green the HSTs are. "HSTs are predominantly electric powered and therefore emissions from HST operations are considered to be linearly related to energy consumption and the sources used to generate the electricity" [15, p. 606]. Thus, the cleaner the source of energy, the more climate friendly HSTs are.

In France, the main energy source of the TGV trains is nuclear energy. Nuclear power is seen as an important source of energy as it generates electricity in higher amounts than any other clean energy source. However, nuclear power carries high risks in terms of radioactive elements produced during the energy production process and the problems of radioactive wastes; thus, great attention needs to be paid in planning these systems especially in places where geographical hazards are more likely to happen.

As discussed earlier, the modal shift from cars and aircrafts to HSTs as a result of increased connectivity between major cities is known to be beneficial for the environment. The CO_2 emissions per passenger generated by HSTs are significantly lower than the emissions from aircraft [16]. While the cleanest aircraft produces 56.7 kg of CO2 per passenger between London and Paris, the HST produces 14.4 kg per passenger on the same route (ibid. 80). Thus, it is quite common to speak about CO_2 benefits as a result of HST operations. However, it is not possible to achieve reductions in CO_2 emissions unless HSTs fully replace road and air traffic. According to Givoni and Banister [14] London and Paris route has not experienced such a decline in air traffic after the introduction of HST services. On the contrary, carriers like British Airways and Air-France continue to offer frequent services on this route even though the HST operator captures over 70% of the market [14, p. 6].

Givoni and Banister [14] mention two conditions that are necessary for the HST to reduce air transport impact on the environment. "First, the HST service must be integrated with other aircraft services provided by the airline and thus it must complement it rather than compete with it. Second, the airline (or airport) should not use the freed runway slots for other flights. With the development of the HST network and successful integration of air and rail services (for example, at CDG) the first condition is now being met at a number of airports" [14, p. 7]. However, unless new regulations are introduced to the airway companies, the second necessary condition is unlikely to happen.

Future Directions: Future of the High-Speed Train

The fast development of the modern HST in the last 50 years and the construction of high-speed lines by a growing number of countries all around the world every year show that the HSTs are regarded as crucial elements for the future of the transport systems.

First of all, the HST development needs to be seen as a network development integrated both within and in-between countries rather than single lines connecting major cities, which is often the case. Thus, integration of HSTs with other transport modes is crucial. As discussed earlier, the development of HST lines in Europe under the Trans-European Network is an important step toward achieving a more inclusive network in Europe. However, further development is needed especially in the Eastern European countries. This can be achieved not only in Europe but also in other parts of the world by "adopting a HST technology where the trains can also run on the conventional network (the TGV model) and by designing the HST stations to allow fast and seamless transfer between HST services and other (national and regional) services" [14, p. 17]. Moreover, a global integration of the whole system can be achieved by possible connectivity of HSTs all around the world via airports.

Second, the uneven distribution of HST network is an important obstacle for the regional development and equality. Although it is not possible to build HST lines connecting every single city due to the high construction and operation costs, it is an important task for planners, engineers, and authorities to make the HST service available to the wider community. Thus the question, as the UIC [34] suggests will be "how to reduce the door-to-door travel time when serving large built-up areas?" [34, p. 1]). In this sense, the use of existing rail systems and their integration with the HST lines is crucial.

Third, as discussed earlier, HSTs can help shifting the demand from other modes of transport to rail as they offer reductions in travel time. This might result in reducing the impacts of transport on the environment. However, there is evidence that the HSTs can create new demand [15]. In this case, "the introduction of HST services might be counterproductive to sustainable transport, even when it promotes rail use" [14, p. 6]. In order to overcome this problem there is a need to ensure the development of HST result in modal shift (e.g., from aircraft) and not generation of new travel. Moreover, as integration is a crucial element in creating sustainable transport, technical incompatibilities need to be solved to integrate the systems fully. Differences in electrical power, safety requirements, track belts, loading gauge sizes are the main technical problems to overcome [20]. Technical standardization has been a major issue for the EU. In order to improve the interlinking and interoperability of the national rail networks as well as access to them, rail interoperability directives have been published [11].

Finally, HSTs can be seen as locomotives of the regional development and can regenerate the economy in regions where the socio-economic deprivation has started as a result of lack of public services and private investment. However, they might also create spatial imbalance and social inequalities. Therefore, introduction of HSTs need to be evaluated within the regional and national socio-economic plans.

In conclusion, High-Speed Train can be considered an important element of sustainable development as it can increase the social integration, act as a catalyst for the economic development and provide greener transport with less energy consumption and high passenger volumes only if the conditions that are discussed in this text are met. Only then it would be possible to discuss the benefits of HSTs not only for a sustainable transport system but also for sustainable local, regional, and national development by all means. It would then be possible to state that the high-speed train marks the second age of the railways.

Bibliography

Primary Literature

1. Banister D (2008) The sustainable mobility paradigm. Transp Policy 15:73–80
2. Banister D (2011) The triology of distance, space and time. J Transp Geogr 19:950–959
3. Banister D, Berechman J (2000) Transport investment and economic development. UCL Press, London
4. Bradsher K (2010) China is eager to bring high-speed rail expertise to the US. NY Times, 7 April 2010
5. Cervero R (2002) Induced travel demand: research design, empirical evidence, and normative policies. J Plan Lit 17(1):3–20
6. Cooper M (2011) Administration pitches big rail projects. NY Times, 8 Feb 2011. www.nytimes.com/2011/02/09/us/09rail.html?ref=highspeedrailprojects
7. Deutche Bahn (2011) The ICE. Deutsche Bahn's fastest train. http://www.bahn.com/i/view/GBR/en/trains/overview/ice.shtml. Accessed 10 Apr 2011
8. Ewing R, Lichtenstein A (2002) Induced traffic and induced development: a literature review, Prepared for the Chester County Planning Commission
9. Ellwanger G, Wilckens M (1994) High speed for Europe. JPN Railway Transp Rev 3:17–25
10. European Commission (2001) White paper, European transport policy for 2010: time to decide. www.ec.europa.eu. Accessed 28 May 2011
11. European Commission (2004) M1 directive of the European parliament and of the council of 19 March 2001 on the interoperability of the conventional rail. www.eur-lex.europa.eu. Accessed 28 May 2011
12. European Commission (2010) EU energy and transport in figures statistical pocketbook 2010. www.ec.europa.eu. Accessed 14 May 2011
13. European Commission (2011) Trans-European networks. www.ec.europa.eu/ten/index. Accessed 12 May 2011
14. Givoni M, Banister D (2008) Reinventing the wheel – planning the rail network to meet mobility needs of the 21st century. TSU working paper, Ref. 1036
15. Givoni M (2006) The development and impact of the modern high speed train. Transp Rev 26(5):593–611
16. Givoni M, Brand C, Watkiss P (2009) Are railways 'climate friendly'? Built Environ 35(1):70–86
17. Givoni M, Rietveld P (2010) Access to rail in urban areas – examination of the number of stations. TSU working paper, Ref. 1044
18. Graham S, Marvin S (2001) Splintering urbanism. Routledge, London
19. Hall P, Banister D (1993) The second railway age. Built Environ 19(3/4):157–162
20. Hall P (2009) Magic carpets and seamless webs: opportunities and constraints for high-speed trains in Europe. Built Environ 35(1):59–69
21. Harvey D (1990) The condition of postmodernity: an enquiry into the origins of cultural change. Blackwell Publishing, Oxford
22. HS1 (2011) Key facts. http://www.highspeed1.com/about/facts/. Accessed 12 May 2011
23. JR-Central (2011) Central Japan Railway Company, about the Shinkansen, safety. http://english.jr-central.co.jp/about/safety.html. Accessed 11 Apr 2011
24. JR-Central (2011) Central Japan Railway Company, company profile. http://english.jr-central.co.jp/company/company/about/history.html. Accessed 11 Apr 2011
25. JR-Central (2011) Central Japan Railway Company, FY2011 key measures and capital investment. Report can be accessed on http://english.jr-central.co.jp/index.html. Accessed 11 Apr 2011
26. Martínez Sánchez-Mateos HS, Givoni M (2009). The accessibility impact of a new high-speed rail line in the UK: a preliminary analysis of winners and losers. TSU working paper series, Ref. 1041
27. Menendez JM, Guirao B, Coronado JM, Rivas A, Rodriguezi FJ, Ribalaygua C, Urefia JM (2002) New high-speed rail lines and small cities: locating the station. In: Brebbia CA, Martin-Duque JF, Wadhwa LC (eds) The sustainable city II. WIT Press, UK

28. Railisa (2011) The UIC statistics database. www.railisa.tsf.it/railisa/ password required and can be obtained from the UIC website
29. RTRI (2011) Japan Railway Technical Research Institute, Overview of MAGLEV R&D. http://www.rtri.or.jp/rd/division/rd77/yamanashi/english/maglev_frame_E.html. Accessed 18 Nov 2011
30. SNCF (2010) Profile and key figures. http://medias.sncf.com/resources/en_EN/medias/MD0006_20100210/file_pdf.pdf. Accessed 11 Apr 2011
31. SNCF (2011) Legal framework. http://www.sncf.com/en_EN/html/media/CH0012-Suppliers/BR1051-Important-laws/MD0005_20100804-Read-article.html
32. Spiekermann K, Wegener M (2008) The Shrinking continent: accessibility, competiveness, and cohesion. In: Faludi A (ed) European spatial research and planning. Lincoln Institute, Cambridge, MA, pp 115–139
33. Tentea (2011) TEN-T projects, Trans-European Transport Network Executive Agency. www.tentea.ec.europa.eu/en/ten-t_projects/. Accessed 11 May 2011
34. UIC (2010) High speed rail: fast track to sustainability. www.uic.org/IMG/pdf/20101124_uic_brochure_high_speed.pdf. Accessed 11 Apr 2011
35. UIC (2011) High speed principles and advantages. http://www.uic.org/spip.php?article443#outil_sommaire_0. 11 Apr 2011
36. UIC (2011b) High speed lines in the world – UIC High Speed Department. http://www.uic.org/spip.php?article573. Accessed 11 Apr 2011
37. Vickerman R (1997) High-speed rail in Europe: experience and issues for future development. Ann Regional Sci 31:21–38
38. European Commission (1996) Council Directive 96/48/EC of 23 July 1996 on the interoperability of the trans-European high-speed rail system. Official Journal of the European Communities, 17 Sep 1996. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:1996:235:0006:0024:EN:PDF. Accessed 22 Aug 2011
39. Taniguchi M (1993) The Japanese magnetic levitation trains. Built Environ19(3/4):235–243
40. Campos J, de Rus G (2009) Some stylized facts about high-speed rail: A review of HSR experiences around the world. Transp Policy 16:19–28
41. UIC (2005) High Speed Rail's leading asset for customers and society. UIC Publications, Paris. Accessible at: /www.uic.asso.fr cited in Campos J, de Rus G (2009)
42. Ampe F (1995) Technopole development in Euralille. In: Banister D (ed) Transport and urban development. E & F.N. Spon, London, pp 128–135

Books and Reviews

Ampe F (1996) Technopole development in Euralille. In: Banister D (ed) Transport and urban development. E & F.N. Spon, London

Banister D, Goodwin MT (2011) Quantification of the non-transport benefits resulting from rail investment. J Transp Geogr 19(2):212–223

Brons M, Givoni M, Rietveld P (2009) Access to railway stations and its potential in increasing rail use. Transport Res A 43:136–149

Campos J, Gagnepain P (2009) Measuring the intermodal effects of high-speed rail. In: Rus G (ed) Economic analysis of high speed rail in Europe. BBVA Foundation, Madrid

Coto-Milla´n P, Inglada V (2004) Social benefit of investment projects: the case for high-speed rail. In: Coto-Millan P (ed) Essays on microeconomics and industrial organisation. Springer, Berlin

De Rus G, Nombela G (2007) Is the investment in high speed rail socially profitable? J Transp Econ Policy 41(1):3–23

Department for Transport (2010) High speed rail London to the West Midlands and beyond – A report to the government by high speed two limited. www.dft.gov.uk

Garmendia M, Ureña JM, Ribalaygua C, Leal J, Coronado JM (2008) Urban residential development in isolated small cities that are partially integrated in metropolitan areas by high speed train. Eur Urban Reg Stud 15:249–264

Givoni M, Banister D (2006) Airline and railway integration. Transp Policy 13:386–397

Givoni M (2007) Environmental benefits from mode substitution - comparison of the environmental impact from aircraft and high-speed train operation. Int J Sust Trans 1(4):209–230

Givoni M, Holvad T (2009) The prospects for European railways: is the second railway age still here or yet to begin? Built Environ 35(1):5–10

Givoni M, Rietveld P (2007) The access journey to the railway station and its role in passengers' satisfaction with rail travel. Transp Policy 14:357–365

González-Savignat M (2004) Will the high-speed train compete against the private vehicle? Transp Rev 24(3):293–316

Hall P (1999) High-speed train and air: competitive or complementary? Paper presented at the Lake Arrowhead transportation symposium, 25 Oct 1999

Holvad T (2009) Review of railway policy reforms in Europe. Built Environ 35(1):24–42

Mokhtarian PL, Salomon I (2001) How derived is the demand for travel? Some conceptual and measurement considerations. Transport Res A 35:695–719

Murikami J, Cervero R (2010) California high-speed rail and economic development: postindustrial market profiles and public policy implications, symposium, Berkeley, 2–3 Dec 2010

Norman C, Vickerman R (1999) Local and regional implications of trans-European transport networks: the channel tunnel rail link. EnvironPlann A 31:705–718

Preston J (2009) Trends in European railways over the last two decades. Built Environ 35(1):11–23

Rietveld P (2002) Why railway passengers are more polluting in the peak than in the off-peak; environmental effects of capacity management by railway companies under conditions of fluctuating demand. Transport Res D 7(5):347–356

Sánchez-Borràs M, Nash C, Abrantes P, López-Pita A (2010) Rail access charges and the competitiveness of high speed trains. Transp Policy 17:102–109

Sasaki K, Ohashi T, Ando A (1997) High-speed rail transit impact on regional systems: does the Shinkansen contribute to dispersion? Ann Regional Sci 31:77–98

Steer Davies Gleave (2004) High speed rail: international comparisons. Final report for the commission for integrated transport (February). http://cfit.independent.gov.uk/pubs/2004/hsr/research/pdf/hsr.pdf

Takagi R (2005) High speed railways: the last 10 years. JPN Railway Transp Rev 40:4–7

Ureña JM, Menerault P, Garmendia M (2009) The high-speed rail challenge for big intermediate cities: a national regional and local perspective. Cities 26:266–279

Whitelegg J, Hultén S, Flink T (1993) High speed trains: fast tracks to the future? Leading Edge Press, Hawes

High-Occupancy Vehicle and Toll Lanes

KATHERINE F. TURNBULL
Department of Landscape Architecture & Urban Planning, Texas Transportation Institute, College of Architecture – The Texas A&M University System, College Station, TX, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Types of HOV Lanes
Access Treatments
Allowed Vehicles and Vehicle-Occupancy Requirements
Operating Hours
Supporting Facilities and Services
Planning
Design
Enforcement
Public Information and Marketing
Case Studies
Future Directions
Bibliography

Glossary

Buses Vehicles designed to carry multiple passengers. Depending on the size, buses may carry between 12 and 100 passengers.

Carpools Vehicles carrying two or more persons.

High-occupancy toll (HOT) lanes HOV lanes that also allow single-occupant vehicle use through payment of a toll.

High-occupancy vehicle (HOV) lanes Roadway lanes designed and operated for buses, vanpools, and carpools at all times of the day or during specific time periods.

Hybrid vehicles Vehicles which operate under a combination of gasoline and electric power.

Low-emission and energy-efficient vehicles Vehicles which typically use non-gasoline power sources and do not emit high levels of pollutants.

Vanpools Vans organized to carry commuters to and from work.

Definition of the Subject

Traffic congestion continues to be a significant issue in metropolitan areas throughout the country. Transportation agencies at the federal, state, metropolitan, and local levels are using a variety of techniques and approaches to improve traffic flow, enhance mobility, and provide travel options.

High-occupancy vehicle (HOV) and high-occupancy toll (HOT) lanes are being used in some metropolitan areas to ease traffic congestion and enhance mobility. HOV lanes provide travel-time savings and improved trip-time reliability by encouraging travelers to change from driving alone to carpooling, vanpooling, or riding the bus. HOV/HOT lanes expand the allowed user groups to include solo drivers or lower-occupant vehicles, who can access the lanes by paying a fee.

The use of HOV/HOT lanes has evolved since the late 1960s. Today, HOV lanes and HOV/HOT lanes are in operation in 32 metropolitan areas in North America. These facilities represent an important element of the transportation system in these areas, enhancing mobility, managing congestion, and providing travel options.

Introduction

The development and operation of HOV/HOT facilities in the USA have evolved over the past 42 years. The opening of the bus-only lane on the Shirley Highway (I-395) in northern Virginia/Washington, D.C., in 1969, the contraflow bus lane on the approach to New York-New Jersey's Lincoln Tunnel in 1970, and the El Monte Busway in Los Angeles in 1973 represents the first freeway HOV applications in the country.

Many of the initial HOV lanes were bus-only applications or allowed buses and vanpools. In an effort to maximize use, carpools became the dominant use group on most projects during the 1970s and 1980s. The vehicle-occupancy requirements for carpools have evolved over time. A three-person per vehicle (3+)-occupancy level was initially used on many projects, but most current facilities use a two-person per vehicle (2+) carpool designation.

In the 1990s, a few areas began experimenting with value pricing projects, often called HOT lanes, which were allowed under the Value Pricing program in previous federal legislation. HOV/HOT lanes expand the allowable user groups to include single-occupant or lower-occupant vehicles for a fee, while maintaining free travel to qualifying HOVs. In addition, the introduction of electronic toll collection (ETC) and other advanced technologies provided greater opportunities for pricing applications on HOV lanes. HOV/HOT lanes enhance mobility and travel options in congested corridors.

The Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU) provides operating agencies with additional flexibility by allowing exempt user groups to use HOV lanes when available capacity exists on an HOV lane. Tolling single-occupant vehicles (SOVs) represents one of the exempt user groups. Low-emission and energy-efficient vehicles, including some types of hybrid vehicles, are allowed to use HOV/HOT lanes in some states.

Today there are some 130 HOV freeway projects in the 32 metropolitan areas in North America. Further, there are12 HOV/HOT lanes in operation in eight metropolitan areas, and more projects in the planning and development stage. Further, 10 states allow low-emission and energy-efficient vehicles to use HOV/HOT lanes.

HOV/HOT lanes involved different designs and operating strategies. Some facilities, such as the I-25 Express Lanes in Denver, are barrier-separated facilities located in the center median of the freeway, with limited access points. Other HOV/HOT lanes are separated from the adjacent general-purpose lanes only by normal paint striping. Access and egress to these facilities may be limited to specific sections or unlimited. The I-394 HOV/HOT lanes in Minneapolis include sections of both types of treatments.

Some facilities are reserved for HOV/HOT vehicles at all times (24/7). Other facilities are reserved for HOV/HOT vehicles only during the morning and afternoon peak periods and are open to all traffic at other times of the day.

Many HOV/HOT projects throughout the country carry more people than the adjacent freeway lane. In addition, they provide trip-time savings and trip-time reliability while enhancing mobility. Further, the revenues from the HOV/HOT lanes in some areas are used to support transit services, improve overall operations, and maintain the facilities.

This article provides an overview of key aspects related to HOV/HOT facilities. The types of HOV facilities and supporting elements are described next. Factors to consider in planning, designing, marketing, and enforcing HOV/HOT lanes are then described. Case studies of the HOV/HOT lanes in Denver, Houston, and Minneapolis are presented. The article concludes with a discussion of future directions for HOV/HOT lanes.

Types of HOV Lanes

There are four general types of HOV lanes or busways in separate rights-of-way – exclusive freeway HOV lanes, concurrent flow freeway HOV lanes, contraflow freeway HOV lanes, and arterial HOV lanes [1–4].

Busways in separate rights-of-way are special lanes in separate rights-of-way reserved for buses only during all operating hours. Busways are in operation in Pittsburgh, Miami, Minneapolis-St. Paul, Boston, and Ottawa, Canada.

Exclusive HOV lanes are separated from general-purpose lanes by barriers or wide painted buffers. Exclusive lanes may be reversible. This approach is more expensive due to separation, but easier to enforce and safety enhanced.

Concurrent flow HOV lanes are lanes in the same direction of travel as the general-purpose lanes separated by paint striping and in some cases a 2–4 ft buffer. Concurrent flow HOV lanes are the most common type of HOV facility found in the USA.

Contraflow HOV lanes use a lane in the off-peak direction of travel for HOVs in the peak direction of travel. They are typically separated from the general-purpose lanes by pylons or moveable barriers. These types of lanes are appropriate when there is available capacity in the off-peak direction of travel. They are typically peak-only operation, and in some cases, only the morning peak period. There are only five contraflow lanes in operation: three in New York/New Jersey, one in Boston, and one in Dallas.

HOV lanes are also in operation on some arterial streets. Most of these are lanes reserved for buses only. Most are also found in downtown areas or on streets approaching downtown areas. There are a few arterial street HOV lanes open to carpools and vanpools.

Access Treatments

A number of different access alternatives may be used with HOV/HOT lanes. These alternatives include continuous access, slip ramp and limited access, direct access, and freeway-to-freeway HOV connections [1].

Continuous access is common with many concurrent flow HOV lanes. Vehicles may move into and out of the HOV lanes at any point. This approach offers more flexibility, but is more difficult to enforce and may cause safety concerns.

Slip ramps or limited access may be used with concurrent flow or exclusive lanes. This approach provides specific locations for access. This approach also limits flexibility for users, but enhanced safety and enforcement.

Direct access uses flyover ramps and other ramp designs to provide direct access to HOV lanes for park-and-ride lots and transit centers. This approach has higher capital costs, but provides additional travel-time savings and is easier to enforce and enhances safety.

A few areas, such as Los Angeles, have connectors from HOV lanes on one freeway to HOV lanes on another freeway. These connectors have high capital costs, but can provide significant travel-time savings to HOVs.

Allowed Vehicles and Vehicle-Occupancy Requirements

Some HOV facilities allow only buses. These include busways and a few freeway lanes, such as the I-495 contraflow lane in New York/New Jersey. A few HOV lanes are restricted to only buses and vanpools. These include the contraflow lanes on the Long Island Expressway and The Gowanus Expressway in New York City. Most HOV lanes allow buses, vanpools, and carpools. SAFETEA-LU also includes requirements for eligible vehicles, and allows operating agencies to allow other types of exempt vehicles, including low-emission and energy-efficient vehicles and some types of hybrid vehicles.

There are three commonly used vehicle-occupancy requirements with HOV lanes. These requirements are three people per vehicle (3+), two people per vehicle (2+), and variable requirements (2+/3+). A three-person per vehicle (3+) requirement is used on the Shirley Highway in northern Virginia and the approach to the Bay Bridge in Oakland. Most HOV lanes that allow carpools use a 2+ requirement. The variable requirements (2+/3+) are used on the El Monte busway in Los Angeles and US 290 HOV lanes in Houston.

Operating Hours

Four different hours of operation commonly used with HOV facilities. Some HOV lanes operated 24 h a day/7 days a week (24/7). Peak-period only – typically morning and afternoon (in the range of 6:30–9:00 a.m. and 3:00–6:00 p.m.) – is found on many HOV lanes. Peak-period only/peak direction, typically morning peak-period only, is found on some HOV lanes. The Houston HOV system is one example of the use of extended hours with operation hours from 5:00 a.m. to 12:00 p.m. in the inbound direction and from 2:00 to 9:00 p.m. in the outbound direction.

Supporting Facilities and Services

Supporting facilities and programs are typically found with HOV lanes to help encourage travelers to change their mode. Examples of supporting facilities and services include new and expanded bus services, park-and-ride and park-and-pool facilities, and transit stations. Newer or expanded express bus routes are often provided with HOV/HOT lanes. Rideshare programs, vanpool programs, and employer-based programs are also key elements of many successful HOV/HOT projects.

Planning

Planning for transportation improvements, including HOV/HOT facilities, usually occurs at different levels. A broad regional planning effort is often undertaken first. This level focuses on the general needs, issues, and opportunities throughout a metropolitan area. The outcome of this process is a long-range regional plan that identifies the general types of facilities anticipated in the major travel corridors and areas. Regional plans may not define the exact type of treatment or design, however.

These more detailed analyses are conducted in the planning process at the corridor and facility level. Planning at this stage is much more detailed, focusing on preliminary alternative design treatments, access options, vehicle-occupancy levels, and other issues. For example, a regional plan may identify the need for an HOV/HOT facility in a specific corridor, but the type of HOV/HOT lane and the preliminary design would be examined in later planning phases.

Planning for HOV/HOT facilities can be thought of as a cascading or funneling process. Planning at the regional scale focuses on identifying an overall system plan and the general types of facilities that may be appropriate. Planning at the corridor and facility level is more detailed, resulting in projects that can be taken forward into the design stage. Different models and analytical tools and techniques are needed at various levels. Sketch planning techniques, travel forecasting models, and more detailed models and methods may be used.

Design

The design of an HOV/HOT lane is similar to the design of a general-purpose traffic lane. There are special elements that should be considered with the design of an HOV/HOT lane, however. Elements that may influence design considerations of an HOV/HOT lane include the separation between the HOV/HOT lane and the general-purpose freeway lanes, access points and the types of access, the toll payment method and the method for HOV declaration or observation, the number and nature of enforcement areas, and signs and markings.

The design of HOV/HOT lanes should follow the standards recommended by the American Association of State Highway and Transportation Officials (AASHTO), the Institute of Transportation Engineers (ITE), and the Transportation Research Board (TRB). The design for different types of HOV facilities have been examined by these organizations and other groups, and are documented in different reports [5–7]. In addition, Texas [8], California [9], Washington [10], and other states have developed design guidelines.

Enforcement

Enforcement is a key component of operating HOV/HOT lanes. The roll of an HOV/HO enforcement program is to ensure that operating requirements, including vehicle-occupancy levels, are maintained to protect HOV travel-time savings, to discourage unauthorized vehicles, and to maintain a safe operating environment. Visible and effective enforcement promotes fairness and maintains the integrity of the HOV facility to help gain acceptance of the project among users and nonusers. Ensuring that tolls are paid is another key element of HOV/HOT enforcement. Key elements of successful enforcement programs are highlighted in this section [1].

First, the appropriate agency must have the legal authority to issue citations for possible violations. The appropriate agency may include the state police, local police, and transit police. Toll authorities involved in HOT projects must also have the legal authority to enforce toll violations.

State legislation or local laws may be needed to ensure that the appropriate agencies and law enforcement personnel have the legal authority to issue citations for various offenses, including vehicle eligibility requirements, vehicle-occupancy requirements, operating regulations (such as entering and exiting an HOV lane at the appropriate places), and not paying tolls. If new user groups (tolled vehicles, low-emission and energy-efficient vehicles) are allowed to use an HOV/HOT lane a check should be made to be sure the legal authority for enforcement exists. Legislative action may be necessary if the legal authority does not exist.

Two elements of HOV enforcement are the type of citation issued for violating the HOV requirements and the level of the fine. Most areas classify occupancy violations and entering and existing the HOV lanes illegally as a moving violation. Some areas include demerits on a driver's record. Fines in different states for violating occupancy requirements and other operating procedures range from $40 to $500 for repeat offenses. Some areas use a graduated scale, with fines increasing for repeat offenses, and some areas add court costs.

Four general enforcement strategies are typically used with HOV facilities. These strategies are routine enforcement, special enforcement, selective enforcement, and peer-enforcement.

Routine enforcement maintains the normal level of law enforcement in the corridor or freeway, with no extra enforcement designated for the HOV lane. Routine enforcement is used with many HOV facilities around the country during some or all operating hours. Advantages of this approach include no additional funding needed and usually some level of enforcement provided. Disadvantages include the potential for a low level of enforcement, monitoring HOV facility not a priority, and a potential for high violation rates.

Special enforcement includes the dedication of additional law enforcement personnel and resources to the HOV facility. Approaches may include assigning a patrol car or motorcycle to an HOV lane, adding extra patrols in a corridor, or locating enforcement personnel along the lane. Special enforcement may be accomplished by reallocating existing personnel or by adding additional personnel during key periods. Advantages include a more visible police presence, lower violation rates, and enhanced public perceptions. A disadvantage is the costs associated with extra personnel and the need for additional funding.

Selective enforcement provides extra enforcement personnel at specific times or in response to certain situations. These situations may include when a new HOV facility is open, in response to high violation rates, when major operation changes are made, or periodically to reinforce occupancy and other requirements. Advantages include higher visibility, positive public perception, and lower violation rates. Disadvantages include the cost, which requires additional funding and may become too predictable.

The final general enforcement strategy is peer-enforcement, which involves peer-regulation by HOV lane users and motorists in the general-purpose lanes by providing a telephone hotline to report potential violators. Peer-enforcement is typically used in addition to other enforcement methods – not as the only strategy. Peer-enforcement provides a telephone number people can call to report the license plate of a vehicle they feel is violating the HOV requirements. Typically, a letter is sent to the vehicle owner with information on the correct use of the HOV/HOT lane. License plates reported more than once typically receive more strongly worded follow-up correspondence and may be monitored by enforcement personnel in the field. The HERO program in Seattle is the best example of a peer-enforcement program.

Specific enforcement techniques include the use of stationary patrols, roving patrols, team patrols, multipurpose patrols, electronic assistance, and citations or warnings by mail. Stationary patrols are law enforcement personnel located at dedicated enforcement areas or other areas. It is important that these areas provide a safe location for enforcement personnel. The dedicated enforcement area is located just before the exit point of some HOV/HOT lanes.

Roving patrols involve the use of law enforcement cars or motorcycles patrolling the length or a portion of an HOV/HOT facility. A safe area to pull vehicles over is still needed with this approach. These enforcement vehicles may travel in the HOV lane or the adjacent general-purpose lanes. A safe area to pull vehicles over and stop is still needed with this approach.

Team patrols use a combination of stationary and roving patrols. Law enforcement personnel maintain radio contact to identify potential violators, which may be pulled over by downstream enforcement personnel. Multiple enforcement points may also be used.

Multipurpose patrols may provide HOV enforcement as one of many duties. These extra patrols may be located in congested corridors to assist with incident management and other functions.

Closed-circuit television cameras (CCTV), infrared cameras, and photographing license plates may be used to assist in enforcing HOV lanes. To date, no technology is commercially available to detect the number of individuals in a vehicle. Public and institutional acceptance of electronic monitoring may be an issue in some areas.

Legal authority is needed for enforcement personnel to issue warnings or citations for violating HOV/HOT requirements by mail. This approach eliminates the need to stop a vehicle. At present, no ticketing by mail of HOV/HOT violations is in effect in the USA.

It is important that information on enforcing HOV/HOT facilities, especially the citation and fines for violating requirements, be communicated to motorists and the public. It is also important that the courts and judicial system is informed of the HOV penalties and the importance of the facilities to help ensure that citations are upheld.

Public Information and Marketing

Public involvement is a key element of the transportation planning and project development process, including planning HOV/HOT lanes. The term relates to proactively involving the public in the discussion of transportation plans and projects. There are specific federal and state requirements associated with public involvement, including those contained in the National Environmental Policy Act (NEPA) and SAFETEA-LU.

Public and stakeholder outreach is a key element in the planning, project development, implementation, and operation phases of HOV/HOT projects. Public and stakeholder outreach is focused on building understanding and support for the project with targeted groups. These targeted groups might include members of the state legislature, local elected officials, business organizations, neighborhood groups, and other organizations. Examples of possible techniques include collaborative task forces and advisory committees. Public information represents another key element in the planning, project development, implementation, and operation phases of HOV/HOT projects. Public information focuses on building awareness, understanding, and support for an HOV/HOT project among travelers and the public. Potential techniques include printed and electronic newsletters, press releases, Internet sites, presentations, and meetings and workshops.

Market research is frequently used in the planning and project development stages of HOV/HOT projects. Market research helps test potential concepts, projects, and messages. Examples of market research techniques include focus groups, interviews, and mail, telephone, and Internet surveys. Examples of market research techniques used with HOT projects include focus groups in San Diego and Minneapolis and surveys of bus riders, carpoolers, and HOT lane users in San Diego.

Promotion and marketing is used during the implementation and operation phases of HOV/HOT projects. These efforts and techniques focus on promoting the initial and ongoing use of the HOV/HOT lanes. Possible marketing techniques include paid television and radio advertisements, television and radio public service announcements, paid advertisements in neighborhood and city or regional newspapers, the Internet, direct mail, service centers, and billboards.

A number of elements are typically associated with developing a marketing program for an HOV/HOT project. First, the target markets are identified. Examples of target markets include households in the corridor, commuters in the corridor, and major employers in the corridor. The marketing concepts, messages, and branding of the HOV/HOT project may be tested through the use of the different market research techniques described earlier. Based on the results of the market research, the branding and messages are finalized. The appropriate media and approaches, which were described previously, are identified. Media and approaches are matched with available funding and local perceptions to develop the marketing plan to be implemented.

Three phases are typically associated with marketing HOV/HOT projects. These phases are preopening, opening/initial operation, and ongoing operation. The focus of the marketing program during the preopening phase is on building awareness of the requirements for using the HOV/HOT lanes and promoting individuals to purchase transponders. During the opening and initial operation of an HOV/HOT project, marketing efforts continue to focus on explaining the use requirements and selling transponders. The marketing effort during this phase also promotes the use of the HOV/HOT lane by all groups – toll paying SOVs, carpools and vanpools, and bus riders – and converting travelers violating the toll or vehicle-occupancy provisions into users. Ongoing marketing efforts focus on promoting the use of all options, as well as transponder sales.

A variety of approaches may be used to market HOV/HOT projects. These approaches include Internet sites, direct mail, regional and neighborhood newspaper advertisements, radio and television advertising, radio and television earned media or public service announcements, and customer service centers. Brochures, which may be distributed at meetings, by direct mail, or at businesses or service centers, represent another marketing approach used with many existing HOV/HOT projects.

Case Studies

Denver

The I-25 HOV lanes were opened in the mid-1990s. Called the Downtown Express, the HOV lanes are 7 miles in length. The barrier-separated facility includes two lanes with shoulders on each side. Access is provided at both ends of the lane. The lanes operate in the inbound direction toward downtown Denver in the morning and in the outbound direction in the afternoon. A 2+ vehicle-occupancy requirement is used.

Although carpool, vanpool, and bus use of the HOV lanes was good, available capacity existed. Consideration of expanding the eligible user groups to include toll paying solo drivers began in the late 1990s. Enabling legislation was approved in 1999 to allow HOT projects in the state.

Implementation of the Express Lanes included adding ETC and new signs along the facility. Drivers entering the Express Lane self-declare as HOV or toll paying vehicles by using the appropriate access lane. In June 2006, toll paying solo drivers were allowed to use the I-25 HOV lanes. The Express Lanes use ETC, with preset variable pricing by time-of-day. The current fees range from a low of $.50 on Saturdays, Sundays, and off-peak periods to a high of $3.25 during peak times.

Solo drivers must obtain a transponder and maintain an active account to use the Express Lanes. The transponders can also be used on the E-470 and the Northwest Parkway toll facilities.

The development and operation of the HOV lanes and the Express Lanes represent the coordinated efforts of the Colorado Department of Transportation (CDOT) and the Regional Transit District (RTD). The Colorado Tolling Enterprise (CTE), a part of CDOT, assumed operating responsibility when the Express Lanes were initiated. The Colorado State Patrol and the E-470 Tollway Authority are responsible for the enforcement of the I-25 Express Lanes. Video enforcement is used to identify vehicles without valid toll tags.

Traffic counts taken in November 2006, 6 months after the implementation of the HOT project, show toll vehicles accounting for between 28% and 32% of the vehicles using the facility. Of the total 1,043 vehicles using the lanes in the afternoon peak hour, 287, or 24%, were tolled vehicles, and 756, or 73%, were HOVs. In the morning peak hour, tolled vehicles represented 32% of the total 899 vehicles, while HOVs accounted for 68%. The travel time for buses and carpools has remained relatively constant [11].

Houston

The Houston metropolitan region is characterized by low-density development, typical of southwestern cities. In response to ongoing concerns related to traffic congestion on the freeway system, limited available right-of-way, and air quality, an initial 9-mile contraflow demonstration project has evolved over an almost 30-year period into a system that encompasses some 100 miles of HOV lanes, numerous direct access ramps, 28 park-and-ride lots, 4 park-and-pool lots, an extensive network of express bus service, and a managed lanes/HOT lane project. This system provides preferential treatment to buses, vanpools, and carpools in the major freeway corridors.

Planning, designing, operating, and enforcing the HOV system elements has been accomplished through the coordinated efforts of the Texas Department of Transportation (TxDOT) and the Metropolitan Transit Authority of Harris County (METRO). Recently, Harris County and the Harris County Toll Road Authority (HCTRA) have joined this partnership to assist with the development and operation of the managed lanes on I-10 West. These efforts have been coordinated with the Houston-Galveston Area Council (HGAC), the metropolitan planning organization (MPO) for the area.

The HOV lanes are primarily one-lane, reversible, barrier-separated facilities, located in the median of six freeways. A short two-lane, two-direction section exists on the Northwest (US 290) Freeway. A two-way facility, with one lane in each direction of travel, is in operation on the Eastex (US 59) Freeway. A concurrent flow HOV lane is in operation on the Katy Freeway, leading to the reversible lane. The design used for the HOV lanes was influenced by a number of factors, including limited right-of-way in the freeway corridors, providing a safer operating environment using barriers, and the directional splits in the corridors.

The lanes operate in the inbound direction from 5:00 a.m. to 12:00 p.m. and in the outbound direction from 2:00 to 9:00 p.m. The lanes are closed from 12:00 to 2:00 p.m. to reverse operations and are closed to all traffic at other times. A 2+ vehicle-occupancy requirement is used on all the HOV facilities, except the Northwest Freeway. This facility has a 3+ occupancy requirement from 6:45 to 8:00 a.m. and 5:00 to 6:00 p.m., due to congestion occurring at the 2+ level. The Quick Ride value pricing project operates on these two lanes, allowing participating 2+ carpools use of the lane for a $2.00 per trip fee.

A total of 28 park-and-ride lots and four park-and-pool lots are located in the six corridors with HOV lanes. The larger park-and-ride lots have direct access to the HOV lanes and transit stations with passenger amenities. There are spaces for between 900 and 2,500 automobiles at 19 of the lots. The number of parking spaces at lots in each corridor range from slightly over 3,000 to almost 7,500.

The park-and-ride lots have transit stations with covered passenger waiting areas and other amenities. Transit centers without park-and-ride lots or with only small lots are located at strategic transfer points.

Direct access ramps connect the major park-and-ride lots and transit stations to the HOV lanes. These ramps provide travel-time savings for buses using the HOV lanes and enhance the safe operation of both the HOV lanes and the freeways. Use of the direct access ramps is restricted to buses, carpools, and vanpools during operating hours. The ramps are closed during nonoperating periods. Carpools and vanpools can access the ramps and the HOV lanes through the lots. The direct access ramps provide significant travel-time savings for buses and other HOVs. The 1990 opening of the direct access ramp linking the Northwest Freeway Station park-and-ride lot with the Northwest Freeway HOV lane provided travel savings of 14 min for vehicles entering and exiting the HOV lane. Prior to the ramp opening, HOVs had to travel local streets, enter the freeway, and merge across the general-purpose lanes to enter the HOV lane. Use levels increased after the ramp opened.

METRO provides a high level of bus service in each corridor, with frequent trips from the major park-and-ride lots. Over-the-road coaches are operated on many routes, as are articulated buses. Although there is no direct evidence linking increased ridership to use of the coaches, surveys of bus riders indicate support for their use and support for frequent service. The HOV lanes and express bus services are oriented primarily in a radial direction, with downtown Houston as the major destination. The express bus system has evolved over the years, however, providing service to major activity centers such as the Texas Medical Center (TMC), Greenway Plaza, and the Post Oak/Galleria area. More recently, reverse commute services have been added in some corridors, taking advantage of buses in the general-purpose lanes deadheading back to park-and-ride lots.

METRO also provides rideshare services in the Houston area. METRO's RideShare program includes a number of elements to help individuals form carpools and vanpools. Rideshare matching services are available by telephone and online through METRO's Internet site. The METROVan program helps commuters form vanpools and provides vans for their use. METROVan is cosponsored by HGAC, allowing METRO to provide vanpools outside the METRO service area. METRO's corporate RideSponsor program focuses on encouraging employees to commute to work by bus, carpools, or vanpools. The program provides computerized ridematching services, vanpools, and employer outreach. Corporate RideSponsors are eligible for discounted bus passes for their employees.

In 2003, some 212,079 passengers used the HOV lanes on a daily basis. Buses carried 43,225 passengers, vanpools accounted for 2,500 riders, carpools had 74,867 occupants, and 407 motorcycles used the lanes daily. Morning peak hour utilization levels range from approximately 1,000 vehicles on the Katy Freeway HOV lane to 1,551 on the Northwest Freeway HOV lanes.

Corresponding person volumes in the morning peak hour average between 3,424 on the Gulf Freeway HOV lane and 4,836 on the Northwest Freeway HOV lane. The HOV lanes account for 40% of the morning peak hour total person movement on three of the freeways. The AM peak hour is defined as the hour with the highest vehicle volumes. As a result, the peak hour may vary by HOV lane.

The HOV lanes and direct access ramps have significantly increased METRO bus operating speeds. The peak hour bus operating speeds have almost doubled, from 26 to 54 mph, resulting in significant reductions in bus schedule times. Examples of reductions in the morning peak hour schedule time for buses from park-and-ride lots to downtown Houston include from 45 to 24 min from the Addicks park-and-ride lot on the Katy Freeway HOV lane, from 40 to 25 min from the Edgebrook park-and-ride lot on the Gulf Freeway HOV lane, and from 50 to 30 min from the Northwest Freeway Station park-and-ride lot on the Northwest Freeway HOV lane.

The HOV lanes provide travel-time savings for buses, vanpools, and carpools. Morning peak hour travel-time savings range from approximately 2 to 22 min on the different HOV lanes. The Northwest Freeway HOV lane generally provides the largest travel-time savings of about 22 min. The Katy Freeway HOV lane averages between 17 and 20 min, the Northwest Freeway 14 min, and the Gulf Freeway and Southwest Freeway between 4 and 2 min. In addition, the HOV lanes provide more reliable trip times to carpoolers, vanpoolers, and bus riders.

Approximately 32,293 spaces are provided at 28 park-and-ride lots associated with the HOV lanes. An additional 1,377 spaces are located at four park-and-pool lots. METRO buses serve the park-and-ride lots, while the park-and-pool lots provide staging areas for carpools and vanpools. In 2003, the overall occupancy levels at the individual facilities ranged from about 10% at some park-and-pool lots to 100% at well-used park-and-ride lots. From 1980 to 1990, the number of park-and-ride lots doubled from 10 to 20. The number of available spaces increased from 4,070 spaces to 12,626 spaces. Use of the lots grew from 4,070 parked vehicles to 12,626 vehicles. As of 2003, there are 28 park-and-ride lots, with 32,293 spaces. Approximately 54% of the available spaces are used on a daily basis [12].

The travel-time savings and the improved trip-time reliability have influenced commuters to change from driving alone to taking the bus, carpooling, and vanpooling. Periodic surveys of HOV lane users show that between 36% and 45% of current carpoolers formerly drove alone, while 38–46% of bus riders previously drove alone. Surveys conducted in 1988, 1989, and 1990 indicate that the opening of the HOV lanes was very important in their decision to ride a bus for between 54% and 76% of the bus riders using the Houston HOV lanes. Between 22% and 39% of the respondents also indicated that they would not be riding the bus if the HOV lane had not been opened.

The HOV system has resulted in an increase in available vehicle-occupancy (AVO) levels in the corridors with HOV lanes. For example, the morning peak hour AVO increased on the North Freeway from 1.28 in 1978 before the contraflow HOV lane opened to 1.41 in 1996. The morning peak hour AVO increased on the Northwest Freeway from 1.14 in 1987 prior to the opening of the HOV lane to 1.36 in 1996. The 1996 morning AVO for the HOV lanes ranged from 2.6 to 3.65, compared to 1.02 to 1.12 for the general-purpose lanes.

Periodic surveys of HOV lane users and motorists in the general-purpose lanes included questions designed to obtain feedback on the general perception toward the HOV lanes and support for these facilities. Between 40% and 81% of motorists in the general-purpose lanes on freeways with HOV facilities and one freeway without an HOV lane have responded positively to these surveys that the HOV facilities are a good transportation improvement.

A demonstration, called QuickRide, which uses an electronic toll collection system, was implemented at the end of January 1998 on the Katy Freeway. It was later expanded to the Northeast Freeway. Although use levels were modest, the demonstration was successful at allowing an additional user group to use the HOV lanes during the 3+ restricted period. It appeared that many enrollees view having an electronic tag as insurance for the occasional need and opportunity to ensure a quick trip, but cannot use the program on a regular basis.

Plans for expanding the Katy Freeway began in the late 1990s. The 23-mile corridor carries some 280,000 vehicles a day. The existing cross section in the most congested section from SH 6 to I-610 includes three general-purpose lanes in each direction, the HOV lane, and three-lane frontage roads in each direction.

A number of alternatives were examined in the Major Investment Study (MIS), including four special-use lanes in the freeway median. Other options included additional general-purpose lanes and expanding the HOV lanes. The special-use or managed lane option emerged from this study as the preferred alternative. The specific operation of the managed lanes was not defined, but user groups considered including buses, 3+ HOVs, QuickRide HOVs, vanpools, trucks, and long-distance travelers. Other than QuickRide participants, tolling was not considered as an operational strategy. The cross section for the section between SH 6 and I-610 would include a three-lane frontage road and four main lanes in each direction of travel, and the four managed lanes.

During the environmental impact statement (EIS) process, the Harris County Toll Road Authority (HCTRA) expressed interest in participating in the managed lanes portion of the project. The EIS was modified to include tolling. Additional public involvement activities were conducted, along with a more detailed assessment of potential toll-related issues. The constructed facilities include four general-purpose freeway lanes, two managed lanes in each direction of travel. HCRTA helped fund the construction of the facility and operates the managed lanes. Carpools, vanpools, and buses travel for free in the managed lanes, while SOVs can use the lanes by paying toll.

Minneapolis

Two HOV/HOT lanes are currently in operation in the Minneapolis-St. Paul metropolitan area. HOV/HOT lanes are in operation on I-394 to the west of downtown Minneapolis and on I-35W to the south of downtown Minneapolis. Both facilities were initially implemented as HOV lanes and both were expanded to HOV/HOT lanes to maximize and enhance mobility in the corridors.

I-394 was constructed on the alignment of TH 12, a four-lane signalized arterial. An interim HOV lane, called the Sane Lane, was opened in 1985 at the beginning of the construction process. The Sane Lane was implemented for a number of reasons. First, it helped with traffic management during the almost 7-year construction period. Second, it introduced the HOV lane concept to travelers in the corridor and the metropolitan area as a whole. Third, it helped build transit ridership and carpooling for the completed facility.

The I-394 HOV lanes are approximately 11 miles in length and include two different designs. A 3-mile, two-lane, barrier-separated reversible section is located directly to the west of downtown Minneapolis. To the west of this segment are 7 miles of concurrent flow HOV lanes. The reversible lanes provide direct connections with three downtown parking garages, which include bus stops and passenger waiting areas, reduced parking fees for carpoolers, and links to the downtown skyway pedestrian system.

The complete I-394 HOV system was opened in 1992. Although the HOV lanes were well used, averaging between 900 and 1,000 vehicles in the concurrent flow section during the peak hour, interest in considering HOT applications emerged in the early 2000s to help maximize the use of the lanes.

In 2003, state legislation was approved allowing the HOT project on I-394. A task force, comprised of 22 individuals appointed by the Governor, the Lieutenant Governor, and communities in the corridor, was formed to help oversee the project.

MnPASS was undertaken to meet a number of objectives. These objectives included increasing the efficiency of I-394 by increasing the person and vehicle-carrying capabilities of the HOV lanes, maintaining free-flow speeds for transit and carpools in the HOV lanes, and improving highway and transit in the corridor with project revenues. Other objectives focus on developing ETC and advanced technologies to facilitate dynamic pricing and in-vehicle enforcement.

MnPASS implementation activities included restriping the concurrent flow HOV lanes to change from unlimited to limited access, installation of the ETC and electronic enforcement systems, and marketing the sale of MnPASS transponders, which are available online and at the MnPASS Customer Service Center. The MnPASS operating hours on the concurrent flow segment were initially expanded to 24/7. The hours were changed back to the peak hour, peak direction after the 24/7 operation caused traffic congestion in the off-peak travel direction.

The sale of MnPASS transponders and use of the lanes have grown since 2005. As of September 2011, there are 11,890 active I-394 MnPASS accounts and 15,246 transponders. Some 9,000 transponders have been sold. Traffic counts from mid-2006 recorded 1,756 vehicles using the concurrent flow section in the morning peak hour. HOVs accounted for 63% of the traffic, tolled vehicles comprised 32%, and some 5% were toll violators.

MnPASS allows solo drivers to use the HOV lanes on I-394 for a fee. Dynamic pricing is used, with tolls based on the level of congestion in the HOV lanes. The base toll is $.25 and the maximum toll is $8.00.

MnPASS, which was implemented in 2005, represents the first use of tolling in the Minneapolis-St. Paul metropolitan area. MnPASS also represents the first HOV/HOT project in the country on concurrent flow HOV lanes.

Drivers must have a valid MnPASS transponder displayed on the front windshield to use the lanes without meeting the vehicle-occupancy requirement. MnPASS transponders can be purchased online and at the MnPASS Customer Service Center.

MnPASS represents a partnership of the Minnesota Department of Transportation (Mn/DOT), Metro Transit, the Minnesota State Patrol, and local communities.

HOV lanes on I-35W South were extended to HOT lanes between 2009 and 2010 with funding from the US Department of Transportation Urban Partnership Agreement (UPA) program. With the construction of new HOV/HOT lanes through the Crosstown Commons section and the opening of the Priced Dynamic Shoulder Lane (PDSL) approaching downtown, 16 miles of HOV/HOT lanes are in operation. As of September 2011, there were 6,896 active I-35W MnPASS accounts and 7,836 active transponders. There were between 2,570 and 3,209 daily MnPASS users on I-35 W South in September 2011. Approximately 1,000 MnPASS vehicles were using the HOV/HOT lane during the morning peak period in the northbound lane, along with 1,784 carpools/vanpools, and 54 buses [13].

Future Directions

HOV/HOT facilities will continue to play an important role in major metropolitan areas in North America. The evolution from HOVs to HOV/HOT lanes will continue to progress to the broader managed lanes concept. The use of intelligent transportation systems (ITS) and other advanced technologies will allow for additional innovative approaches for maximizing the use of HOV/HOT facilities.

Bibliography

Primary Literature

1. Texas Transportation Institute, Parsons Brinckerhoff Quade and Douglas, Pacific Rim Resources (1998) NCHRP Report 414, HOV systems manual. Transportation Research Board, National Research Council, Washington, DC
2. Institute of Transportation Engineers (1988) The effectiveness of high-occupancy vehicle facilities. Institute of Transportation Engineers, Washington, DC
3. Transit Cooperative Research Program (2006) TCRP Report 95, Chapter 2 HOV facilities, traveler response to transportation system changes. Transportation Research Board, National Research Council, Washington, DC
4. Turnbull K (1992) An assessment of high-occupancy vehicle facilities in North America. Texas Transportation Institute, College Station
5. American Association of State Highway and Transportation Officials (AASHTO) (1992) Guide for the design of high occupancy vehicle facilities. American Association of State Highway and Transportation Officials, Washington, DC
6. American Association of State Highway and Transportation Officials (1989) Roadway design guide. American Association of State Highway and Transportation Officials, Washington, DC
7. Institute of Transportation Engineers Technical Committee 5 C-11(1992) Design features of high occupancy vehicle lanes. Institute of Transportation Engineers, Washington, DC
8. Mounce J, Stokes R (1985) Manual for planning designing and operating transitway facilities in Texas. Publication No. FHWA/TX-851425-2. Texas Department of Highways and Public Transportation, Austin
9. California Department of Transportation (1991) High occupancy vehicle (HOV) guidelines for planning, design and operations. California Department of Transportation, Sacramento
10. Washington State Department of Transportation, Design Manual – Section 1050, High-Occupancy Vehicle Priority Treatment. Washington State Department of Transportation, Olympia
11. Federal Highway Administration (2007) HOV/HOT lanes. Federal Highway Administration, Washington, DC
12. Federal Highway Administration (2003) Houston managed lanes case study: the evolution of the Houston HOV system. Federal Highway Administration, Washington, DC
13. Cofiroute USA (2011) MnPASS express lanes monthly status report. Cofiroute USA, Golden Valley

Books and Reviews

Arnold E Jr (1987) Changes in travel in the Shirley highway corridor 1983–1986. Virginia Transportation Council, Charlottesville

Beihler A (1996) The Pittsburgh busways. In: Conference proceedings, 8th international conference on high-occupancy vehicle systems, Pittsburgh. Transportation Research Board. Unpublished

Betts S, Jacobson L, Rickman T (1983) I-5 HOV lanes: three month report. Washington State Department of Transportation, Olympia

Bonsall J (1993) Development of the Ottawa transitway system. In: Conference proceedings, 6th national conference on high-occupancy vehicle systems, Ottawa, 25–25 Oct 1993.

Transportation Research Record 409. Transportation Research Board, Washington, DC

Bullard D (1991) An assessment of carpool utilization of the Katy high-occupancy vehicle lane and characteristics of Houston's HOV lane users and non-users. Texas Transportation Institute, College Station

California Department of Transportation (1992) Route 55 status information sheet

Christiansen D, Morris D (1990) The status and effectiveness of the Houston transitway system. Texas Transportation Institute, College Station

Crain & Associates (1978) San Bernardino expressway bus evaluation of mixed-mode operations. Southern California Association of Governments

Curitiba's Surface Metro (1993) The urban transport industries report, pp 53–43

Hultgren L, Kawada K, Lawrence S (1998) San Diego's interstate 15 value pricing project [CDROM]. In: 68th ITE annual meeting, Toronto

Institute of Transportation Engineers (1992) More for Less with HOV, The Benefits of High-Occupancy Vehicle Projects

Klusza R (1989) Route 55 three-year status report. California Department of Transportation, Los Angeles

Lancaster A, Lomax T (1987) Proceedings from the second national conference on high-occupancy vehicles lanes and transitways, Houston

McQueen J, Levinsohn D, Waksman R, Miller G (1975) The Shirley highway express-bus-on-freeway demonstration project: final report. U. S. Department of Transportation, Washington, DC

McQueen J, Levinsohn D, Waksman R, Miller G (1985) The Shirley highway express-bus-on-freeway demonstration project: final report. U. S. Department of Transportation, Washington, DC

Metropolitan Washington Council of Governments (1998) 1997 performance of regional high-occupancy vehicle facilities on interstate highways in the Washington region – an analysis of person and vehicle volumes and vehicle travel times. Metropolitan Washington Council of Governments, Washington, DC

Paiewonsky L (1998) A new approach to HOV entry requirements: MassHighway's 3+/limited 2+ sticker program. TRB Pre-print 981081

Pratt R, Copple J (1981) Traveler response to transportation system changes, 2nd edn. U.S. Department of Transportation

SRF, Inc. (1987) Technical memorandum #3 case study: I-394 HOV lane six month data for I-394 Interim HOV lane, Wayzata

SRF, Inc. (1995) I-394 HOV lane case study: final report. Minnesota Department of Transportation, St. Paul

Stockton B, Daniels G, Hall K, Christiansen D (1997) An evaluation of high-occupancy vehicle lanes in Texas, 1996. Texas Transportation Institute, College Station

Stockton B, McFarland F, Ogden M (1998) Feasibility of priority lane pricing of the Katy HOV lane: feasibility assessment. Texas Transportation Institute, College Station

Sullivan E (1998) Impacts and lessons from value pricing on the California state route 91 express lanes [CDROM]. In: 68th ITE annual meeting, Toronto

Sullivan E, Mastako K (1997) Impact assessment for the California route 91 variable-toll express lanes. Transportation Research Board Preprint No. 971046, Washington, DC

Transcore (1997) I-66 congestion management program monitoring project final report. Prepared for the Virginia Department of Transportation

Turnbull K (1988) Proceedings from the 1988 national HOV facilities conference, Minneapolis

Turnbull K (1990) Proceedings from the 1990 HOV facilities conference, Washington, DC, 10–12 Apr 1990. Transportation Research Board Circular 366. Transportation Research Board, Washington, DC

Turnbull K (1991) HOV Facilities – coming of age: fifth national high-occupancy vehicle (HOV) facilities conference. Transportation Research Board Circular 382. Transportation Research Board, Washington, DC

Turnbull K (1991) HOV Facilities – coming of age: national conference on HOV systems, Seattle. Transportation Research Board Circular 384. Transportation Research Board, Washington, DC

Turnbull K (1992) High-occupancy project case studies: historical trends and project experiences. Texas Transportation Institute, College Station

Turnbull K (1992) Moving into the 21st century. In: Conference proceedings, 6th national conference on highway-occupancy vehicle systems, Ottawa. Transportation Research Board Circular 409. Transportation Research Board, Washington, DC

Turnbull K (1992) An assessment of high-occupancy vehicle facilities in North America: executive report. Texas Transportation Institute, College Station

Turnbull K (1996) International HOV facilities. Transportation Research Record 1360, pp 126–137

Turnbull K (1997) HOV, HOT lanes, congestion pricing in the United States. In: Conference proceedings, 25th European transport forum annual meeting

Turnbull K, Capelle D (1998) Development of an HOV systems manual. National Cooperative Highway Research Program, Washington, DC

Turnbull K, Hubbard S (1994) HOV Systems in a new light. Transportation Research Board Circular 442. Transportation Research Board, Washington, DC

Turnbull K, Hubbard S (1994) HOV Systems in a new light. In: Conference proceedings, 7th national conference on high-occupancy vehicle systems, Los Angeles. Transportation Research Circular 442. Transportation Research Board, Washington, DC

Turnbull K, Hall K, Ringrose M (1994) High-occupancy vehicle treatments on toll facilities. Transportation Research Record 1446, pp 14–24

Ulberg C, Farnsworth G, Etchart D, Turnbull K, Henk R, Schrank D (1992) I-5 north high-occupancy vehicle lane 2+ occupancy requirement demonstration evaluation. Washington State Transportation Center, Seattle

Urbitran, Hayden-Wegman (1997) Evaluation of the high-occupancy vehicle (HOV) lanes on the long island expressway: January 1997 HOV lane user and non-user survey. Urbitran, New York

Van Luven H (1995) Implementation of HOV lanes on I-270: lessons learned. In: Proceedings – 7th national conference on high-occupancy vehicle systems: HOV systems in a new light. Transportation Research Record 442, Washington, DC

Virginia Department of Transportation (1997a) I-66 HOV-2 annual report. Commonwealth of Virginia, Richmond

Virginia Department of Transportation (1997b) I-66 HOV-2 demonstration project final report. Commonwealth of Virginia, Richmond

Washington State Department of Transportation (1985) I-5 HOV lanes: 20-month update, Olympia

Hitachi Zosen Inova Technology

Craig Kedrowski, Peter Chromec
Hitachi Zosen Inova U.S.A. LLC, Norcross, GA, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Description of the Elements of the Hitachi Zosen Inova Energy-from-Waste Technology
Future Directions
Bibliography

Glossary

ACS Auxiliary Control Systems are advanced controls used in addition to the basic CCS in order to better control the combustion, specially with large grates or with inhomogeneous waste.

CEMS Continuous Emission Monitoring System: continuous monitoring of stack emission data, as well storing and transmitting data to environmental authorities.

CFD Computational Fluid Dynamics: mathematical model for calculating velocities, temperatures, concentrations, and properties of fluids flowing for example through a combustion chamber.

CCS Combustion Control System: automated system for controlling the combustion process on the grate including frequency, speed, and stroke of the ram feeder and of each grate module (also called GCS), as well as the flow of underfire air to each grate module, overfire and recirculated airflows.

DCS Distributed control system: automated system for controlling and coordinating all the WTE functions.

ECS Equipment control systems are separate control systems dedicated to control individual equipment functions to assure the desired equipment performance.

EfW Energy from waste (synonym to WTE = Waste to Energy).

FGT Flue Gas Treatment.

HHV Higher heating value or gross calorific value (expressed in kJ/kg or Btu/lb): heat released by a material when completely oxidized (reference temperature: 25°C, approx. 77°F, and assuming that H_2O is in liquid form).

HPS High-Pressure Steam: steam from superheater tubes to steam turbine.

LHV Lower heating value (expressed in kJ/kg or Btu/lb): heat released by a material when completely oxidized (reference temperature: 25°C, approx. 77°F, and assuming that H_2O is in vapor form).

LPS Low-pressure steam: usually from extraction of steam turbine, used for internal or external heating tasks.

LRD Load range diagram (or combustion diagram): shows the operating range of a combustion system in terms of fuel input, on one side, and heat input on the other.

O&M Operation and maintenance.

SCC Secondary combustion chamber: usually the vertical chamber above the grate, where gases and ash released from the fuel are completely combusted. Minimum temperature and residence time requirements often refer to the SCC.

SCR Selective catalytic reduction: process for reducing NO_x (nitrogen oxides) usually by means of reaction with an ammonia-containing reagent and in the presence of a catalyst.

SNCR Selective non-catalytic reduction: process for reducing NO_x (nitrogen oxides) usually by means of reaction with an ammonia-containing in the absence of a catalyst.

TOC Total organic carbon: amount of all non-combusted organic carbon in WTE ash; it is also used to express carbon content in MSW.

VOC Volatile organic carbon: concentration of organic compounds in flue gas.

Definition of the Subject and Its Importance

Traditionally, waste has been landfilled – and still is today, when cost is the main criterion for waste management (Fig. 1). “Out-of-sight, out-of-mind” has been the early concept, before science provided a more complete picture of the effects of landfilling on land, water, and air for a long time after a landfill has been filled [1]. The primary reason for these effects is the continuing decomposition of organic waste compounds into methane and other gases that escape to the air, or to water-soluble compounds that dissolve in rainwater percolating through the waste and may contaminate surface and groundwater.

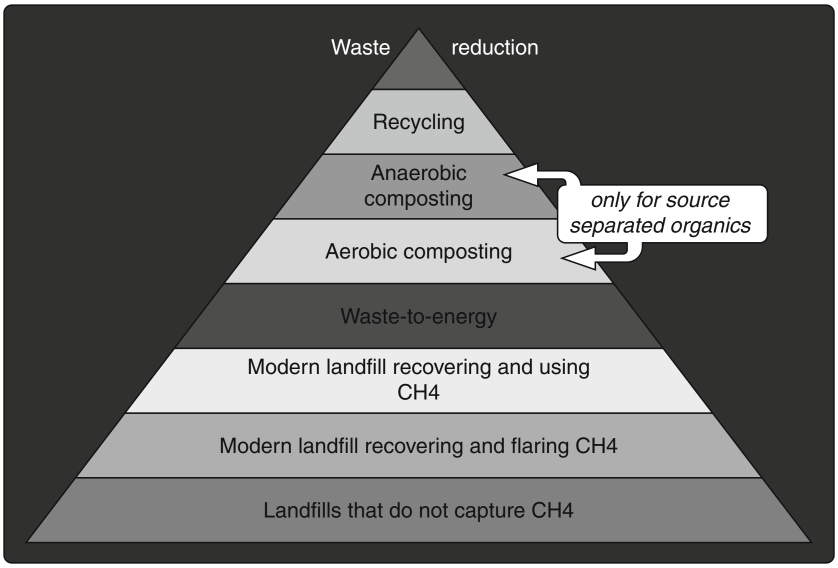
This is why today’s acknowledged hierarchy of waste treatment options considers landfilling as the least desirable option [2]. As shown in Fig. 2, best options are obviously “avoid” and “reuse,” both of which reduce the amount of waste to be treated, and then recycle whatever is reasonably possible from the waste stream. However, the best that has been achieved by recycling and composting is about 70% [3]. This still leaves at least 30% of all waste produced to be treated by waste to energy (also called energy from waste) or landfilling. This entry describes the Hitachi Zosen Inova energy from waste (EfW) technology that recovers the energy contained in the waste and reduces municipal solid waste (MSW) to a maximum 10% of its original volume and to a chemically stable form.

Introduction

The Hitachi Zosen Inova energy-from-waste technology is a mass burn technology which treats post-recycling MSW (i.e., after all reuse and recycling efforts), as collected and without any pretreatment. It consists of a grate-fired steam generator followed by an air pollution control system. Steam is converted to electricity in a conventional steam turbine system. One of the most common configurations is shown in Fig. 3. Collection vehicles dump the waste from the *tipping area* into the *waste pit*. A *waste crane*, manually operated from the *crane control cabin*, or automatically operated by the control system, loads waste from the pit into the *feed hopper*. The waste pit and tipping floor area are ventilated by using the building air as the *primary air intake* for combustion so that no odors can escape to the surroundings. From the *feed hopper*, waste descends through a water-cooled chute and on to the horizontal *feeder table*, over which a hydraulically driven *ram feeder* slides back and forth and pushes waste on to the *reciprocating incineration grate*.

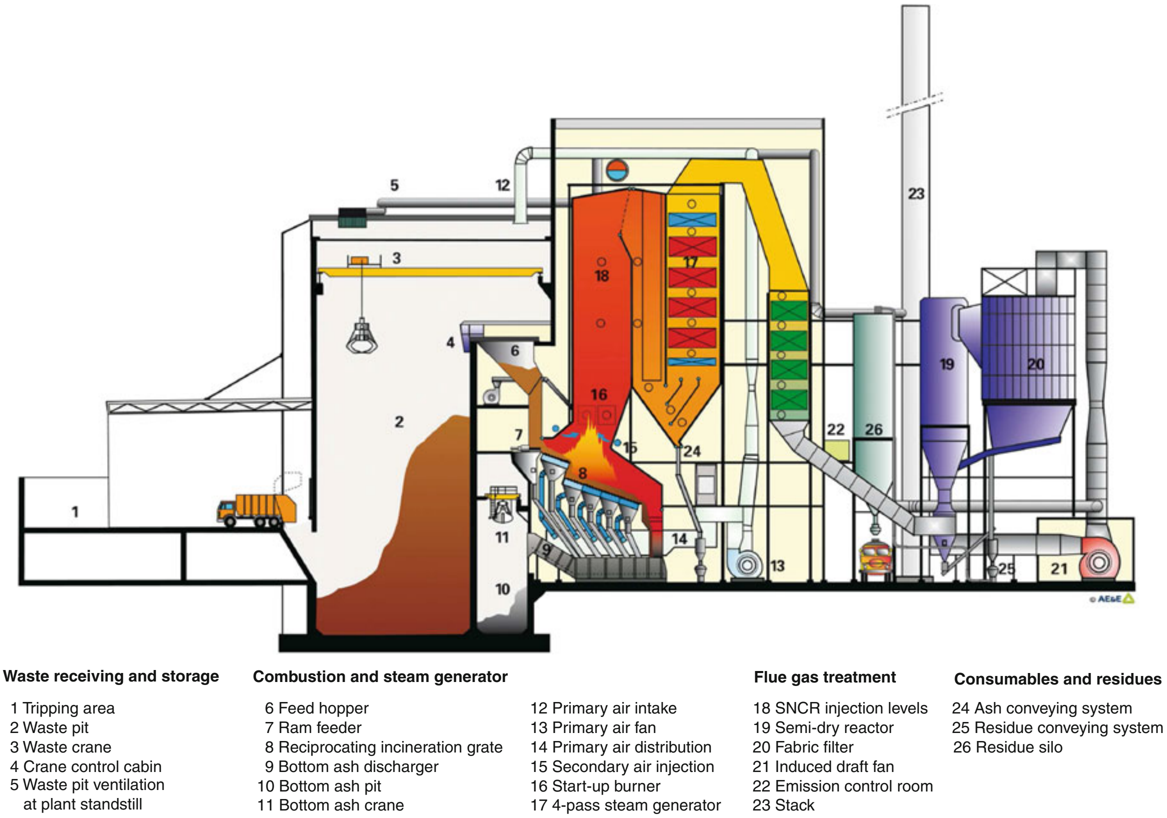
The combustion process is regulated by the combustion control system (CCS). It automatically regulates waste feed, grate movement, and airflow and distribution across the length of the grate to ensure efficient combustion and stable steam generation. Bottom ashes are continuously discharged at the end of the grate, along with siftings falling through the small gaps between the grate blocks, into the *bottom ash discharger* where it is quenched in water.

Hitachi Zosen Inova Technology. Figure 1
From landfilling to a modern energy-from-waste plant


Waste reduction
Recycling
Anaerobic composting
only for source separated organics
Aerobic composting
Waste-to-energy
Modern landfill recovering and using CH4
Modern landfill recovering and flaring CH4
Landfills that do not capture CH4

Hitachi Zosen Inova Technology. Figure 2

Waste treatment hierarchy as per Prof. Themelis [2]



Hitachi Zosen Inova Technology. Figure 3

A typical WTE unit

The ash discharger pushes the bottom ash into the bottom ash pit or alternatively (not shown) into a metal separation system that can recover ferrous and nonferrous metals.

Primary air is injected underneath the grate by the *primary air distribution system*. The *secondary air injection* is right above the grate. At about the same location, a part of the cleaned flue gases is recirculated back into the furnace. Both injections create a swirl in the combustion chamber and assure complete combustion with less excess air and thus a higher efficiency. The *start-up burner* is only needed for start-up and controlled shutdown, as well as to keep temperatures at minimum required levels if wastes with insufficient heat value are treated.

On the way through the *steam generator*, heat from the flue gases is recovered by heat exchangers and used for the production of superheated steam. The heat exchangers consist of membrane walls as well as superheater, evaporator, and economizer tubes. Each of the heat exchangers is designed specifically for the different steam and flue gas temperature ranges and properties. The membrane walls located in the combustion chamber (i.e., first pass) are lined with refractory and/or Inconel cladding to protect against erosion and corrosion. Effective cleaning systems such as sootblowers and water showers are installed to keep fouling of the heat exchanger surfaces within tolerable limits.

The superheated steam produced by the *steam generator* is then piped to a steam turbine / generator which produces electrical energy. A small part of the electricity is used to power the plant. Also, a small part of the low-pressure steam from the turbine is used for heating the plant and preheating combustion air.

The flue gases are cleaned in a multiple-stage flue gas treatment consisting of a selective non-catalytic reduction process (DyNOR®) and a HZ Inova's semidry flue gas treatment process.

The DyNOR® process is based on the principle of selective non-catalytic reduction (SCNR) of nitrogen oxides (NO_x) by injecting ammonia (NH_3) obtained from an aqueous ammonia solution. The *SNCR injection* takes place at the optimum temperature range within the furnace.

After the flue gas has left the boiler, it enters the semidry flue gas treatment process which is designed so that acidic gases are removed by sorption on hydrated lime, while the heavy metal compounds and organic contaminants, such as dioxins and furans, are adsorbed on activated carbon or lignite coke in the same process step. The HZ Inova Semidry FGT process provides intensive contact between the flue gases and the sorbents in a fluidized bed created in the *semidry reactor*. Water is injected to control the temperature in the reactor and to improve the sorption process on the lime. The following *fabric filter* is used for filtering the solid particles from the gases by means of gas-permeable fabric filter bags. A large portion of the filtered solids is recirculated to the semidry reactor in order to improve separation efficiency and sorbent utilization. The rest, generally called fly ash, is transported to the *residue silo* by the *residue conveying system*.

A continuous emission monitoring system (CEMS) is located in the *emissions control room* and monitors continuously the flue gases exiting the air pollution control system. The *induced draft fan* maintains the necessary sub-pressure in the whole plant and assures that the flue gases are conveyed from the furnace to the *stack*.

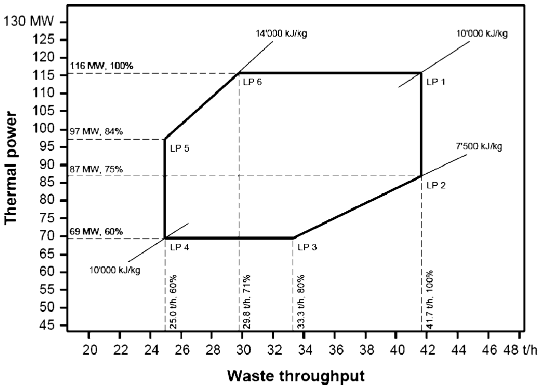
The above parts of the Hitachi Zosen Inova energy-from-waste technology are described in further detail in the following sections.

Description of the Elements of the Hitachi Zosen Inova Energy-from-Waste Technology

HZ Inova Technology Features

General Grate combustion is the most reliable and proven technology for thermal treatment of municipal and similar wastes. HZ Inova has used this technology for over 70 years and has built over 500 units worldwide. Over this period, the technology has continuously been perfected, supplemented with new developments, and adapted to changing requirements. In the following sections, the HZ Inova EfW technology is presented.

Plant Sizes A single combustion train operates within the range of the load range diagram (LRD) shown in Fig. 4. This diagram shows the waste throughput, in tons per hour (t/h), on the horizontal axis and the thermal load, in megawatts (MW), on the vertical axis for different heating values (here lower heat value



Hitachi Zosen Inova Technology. Figure 4

Typical load range diagram of a 1,000 tpd EfW plant

LHV, in kJ/kg) of the waste. The LRD of an EfW plant defines, for a given waste LHV, how much of that waste can be treated in this plant and how much energy will be released during combustion within the operating range of the unit.

HZ Inova EfW units (or "trains") are typically built within the following size range:

Waste throughput (metric tons)	t/h	4–42	Per combustion unit
	t/day	100–1,000	Per combustion unit
	kt/year	33–330	Per combustion unit
Waste			
LHV	kJ/kg	4,500–18,000	
HHV	Btu/lb	2,500–8,400	
Thermal capacity	MW	15–120	Per combustion unit

Recently, with the proven reliability of the HZ Inova EfW technology, ever more single train plants are planned and built. The main question related to single train plants is if waste can be stored or otherwise diverted as well as if it is acceptable to have no heat and/or power supply during service outages. For larger plants, or if redundancy is important, multiple trains can be combined in one EfW plant as needed.

Technology Requirements The waste management market requires for EfW solutions with highest possible flexibility regarding calorific values and composition of wastes to be treated. With changing consumer habits and waste management strategies, including various recycling schemes, properties of feed streams available for an EfW plant can significantly change. At the same time, EfW solutions must meet the highest environmental standards, high thermal efficiency, and be cost effective. The main objectives of the thermal treatment of waste can be summarized as follows:

- Complete inertization and volume reduction of waste
- Destruction of organic pollutants
- Recovery of metals and other useful products
- Concentration of inorganic pollutants in a usable or inert form
- Highest recovery of energy contained in waste for substitution of primary energy
- Cleaning of flue gases to very stringent environmental standards

These objectives translate into the following requirements for a grate combustion system:

- Continuous and controlled waste feed to achieve a constant thermal output despite inhomogeneous waste properties
- Closely controlled supply of combustion air to the individual combustion zones (drying, gasification, ignition, primary combustion, post-combustion) and minimization of infiltration air to ensure good and efficient combustion
- Controlled and adapted movement and stoking of the waste on the grate through the above combustion zones to ensure good burnout
- Efficient and reliable boiler system in order to minimize scaling, corrosion, and erosion in a very difficult environment
- Effective and economic air pollution control system that is adaptable to special local requirements, such as non-attainment zones

Specific main components that make up the Hitachi Zosen Inova EfW technology and fulfill the above requirements are described below in detail.

Waste Pit and Crane

General The purpose of the waste pit and crane is to receive the waste as delivered by the collection or transfer vehicles, store the waste, and also blend it and load it into the waste hoppers so that they are always full.

Waste Pit The waste pit is a large concrete bunker for storage of the waste delivered until it is treated. It serves to uncouple the waste delivery from the waste treatment so that the plant can continue to operate during weekends and holidays when waste delivery is interrupted. Also waste delivery can continue while the plant is shutdown for periodic maintenance.

Another important function of the waste pit is to provide the crane operator with the volume needed to mix the waste. To get the highest capacity and to reduce wear and tear of the equipment, the plant should be operated as stably as possible. This can only be done if the incoming waste, where one truckload could contain completely different waste than the next, is mixed within the waste pit.

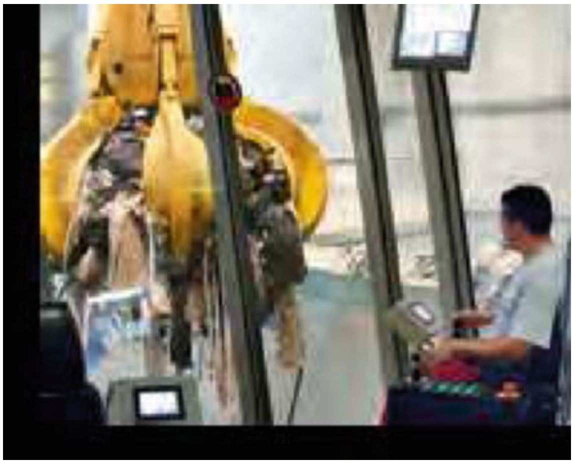
To prevent the combustible waste from igniting in the pit, infrared cameras are installed. When they detect a hot spot, water cannons are used to cool it down.

To prevent the odors in the waste pit from escaping to the environment, either the entire tipping area is enclosed or the dropshafts are equipped with doors. As noted earlier, air is aspirated from the waste pit and used as primary air for combustion. This assures that when the doors are open, air is drawn into the pit and the odor is contained.

Waste Crane A grapple crane (Fig. 5) is used to move the waste in the waste pit for mixing, stacking, and transferring waste to the feed hopper. A single crane can lift loads of several cubic meters of waste in the grapple.

Scales are integrated into the crane in order to weigh every load which is deposited into the feed hopper. This allows the combustion control system to calculate the operating point, monitor trends of waste composition changes, and adapt operating parameters accordingly.

The crane operation can also be automated for certain periods of time so the operator can follow other duties.



Hitachi Zosen Inova Technology. Figure 5
Crane operator loading waste to feed hopper

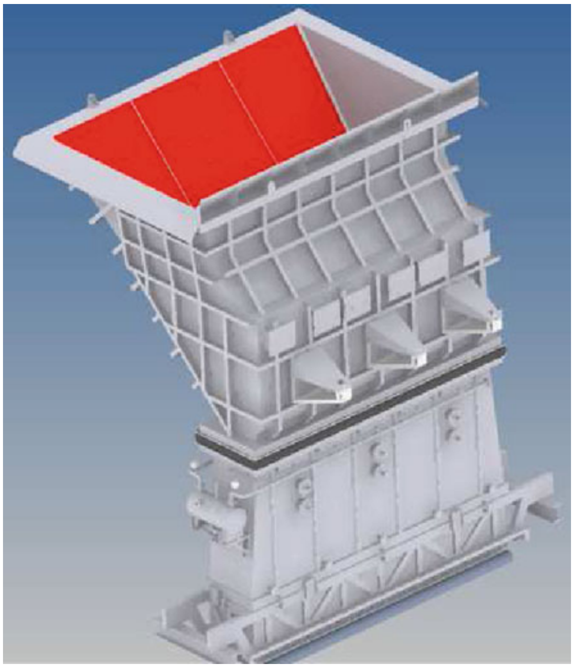
Feed Hopper

General The purpose of the feed hopper/feed chute is to receive waste from the waste pit/bunker via the crane and then to convey the waste onto the ram feeder. It also serves the purpose, by means of the compacted waste column in the chute, of forming a seal between the pit bunker and the combustion chamber, which prevents leakage air into the combustion chamber. Alternatively, the feed chute can be shut off by closing its flap gate.

Description Figure 6 shows the complete feed hopper, consisting of hopper, hopper flap gate, feed chute, and support frame. In order to prevent wear, certain key areas are protected with wear-resistant steel sheet. The slope of the hopper walls and the arrangement of the hopper flap gate are designed to prevent arch buildup and, thereby, ensure continuous delivery of waste to the ram feeder table. The hopper rests on the concrete loading deck above the bunker pit and a sufficiently high protection barrier is provided for personnel safety.

The height of the feed chute facilitates a sufficiently high waste column that seals the combustion chamber from the waste bunker during combustion. The feed chute and the support frame are water-cooled to resist any thermal stress that may occur during operation.

The feed hopper flap gates (Fig. 7) are installed in the bottom of the hopper, one for each grate lane.



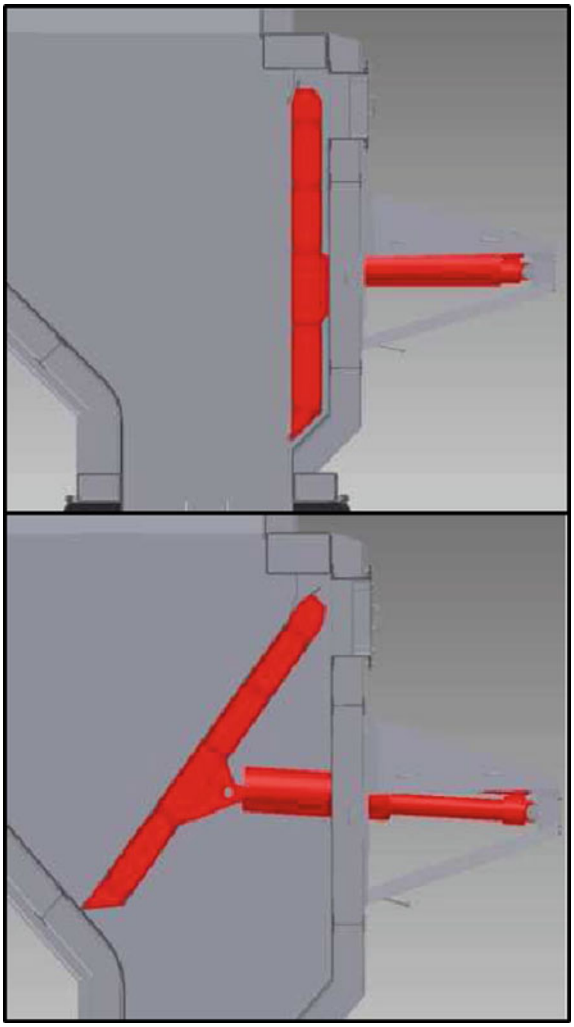
Hitachi Zosen Inova Technology. Figure 6
Feed hopper, chute, and support frame

These gates are hydraulically operated by cylinders powered by the hydraulic station. The hopper flap gates seal the combustion chamber from the bunker prior to waste charging and during system shutdown. An interlock with the combustion chamber temperature ensures that during system start-up the hopper flap gate cannot be opened and waste cannot be charged until the minimum combustion chamber temperature is reached.

In order to monitor the waste height, a reliable, low-maintenance level monitor is installed on the feed chute to provide a low-level alarm. Depending on the requirements for automated operation, one additional level monitor can be installed in the hopper in order to signal to the crane operator to load additional waste. A third-level detector can be installed in order to signal a possible formation of an arch in the feed hopper.

Ram Feeder

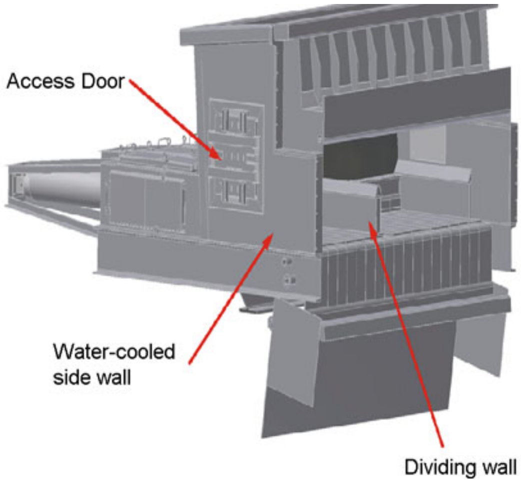
General The purpose of the ram feeder (Fig. 8) is to push the waste from the feed chute and distribute it onto the grate at a controlled rate, as required by the combustion control system.



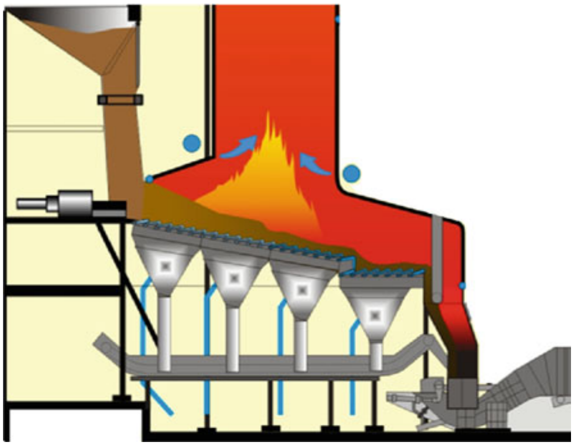
Hitachi Zosen Inova Technology. Figure 7
Feed hopper flap gate

Description The waste descends in the feed chute and falls on to the horizontal feeder table, over which hydraulically driven ram feeders (one for each grate lane) slide. As the feeders slide back, waste falls from the feed chute to the feed table. As the feeders slide forth, waste on the feed table is pushed onto the incineration grate.

The speed and length of the forward stroke can be continuously varied, depending on steam flow demand and waste properties. With each stroke, a defined volumetric quantity of waste is pushed onto the incineration grate. The return stroke is made at full speed. The continuous regulation of the forward stroke allows



Hitachi Zosen Inova Technology. Figure 8
Ram feeder arrangement



Hitachi Zosen Inova Technology. Figure 9
Four-grate modules in a lane

uniform metering and immediate adjustment of waste required on a specific grate row. The strokes are adjusted by the ram feed controller housed in the grate control cabinet located in the immediate vicinity of the waste-charging area. The ram feeders can also be operated manually from this local control cabinet and monitored by the DCS.

Grate

General The purpose of the grate is to incinerate the waste and to ensure continuous and complete combustion resulting in an adequate burnout of the bottom ash to a level of typically less than 3% TOC (total organic carbon).

Hitachi Zosen Inova Technology. Figure 10
Three-lane grate

Description

Overview The basic component of the combustion grate is the grate module. As can be seen in Fig. 9, the four-grate modules are joined together longitudinally to form one-grate "lane." A grate consists of one to four parallel grate lanes. Figure 10 shows a three-lane grate. The width and the number of grate lanes, and therefore the grate size, are determined by the design throughput and the calorific value of the wastes.

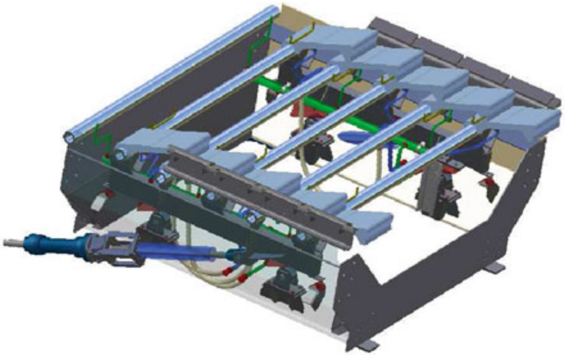
The grate is an inclined forward-feed grate with a slope of 10° to the horizontal. The last grate zone - the burn-out zone - is not inclined, in order to provide for optimal bottom ash quality. The grate, including the riddling hoppers below, is attached to the furnace structure at the feed end by means of a pivot mount which allows it to move. The grate rests on the grate support structure; this allows it to move in response to thermal expansion. The grate cylinders and riddling hoppers move with the grate as it expands and contracts. A boiler-to-grate expansion joint is installed between the grate module and the lower water wall headers.

Each grate module is driven by two hydraulic cylinders. A separate hydraulic control block is provided for each grate element. The stroke acts on the grate carriage located on guide surfaces, thereby cycling the movable grate block rows.

The number, speed, and length of strokes performed by the movable grate block rows can be continuously controlled in response to the process needs and the frequently changing properties of the waste.

Grate Modules Each grate module (Figs. 11 and 12) consists of six grate block rows, alternating between movable and stationary rows. An end piece is bolted onto the last grate module which forms the transition from the grate end to the bottom ash discharge chute.

With the exception of the side plates, the stationary and movable blocks are identical. The grate blocks are supported and mounted on retaining tubes. The block-retaining tubes for the stationary block rows are supported on brackets which rest on the stationary support structure of the grate module. However, the block-retaining tubes for the movable block rows are connected to brackets, which in turn are secured to the supports of the grate carriage that moves in the direction of the waste. The grate carriage is equipped with rollers and guides the movable grate rows. Dual hydraulic cylinders drive the rows back and forth.



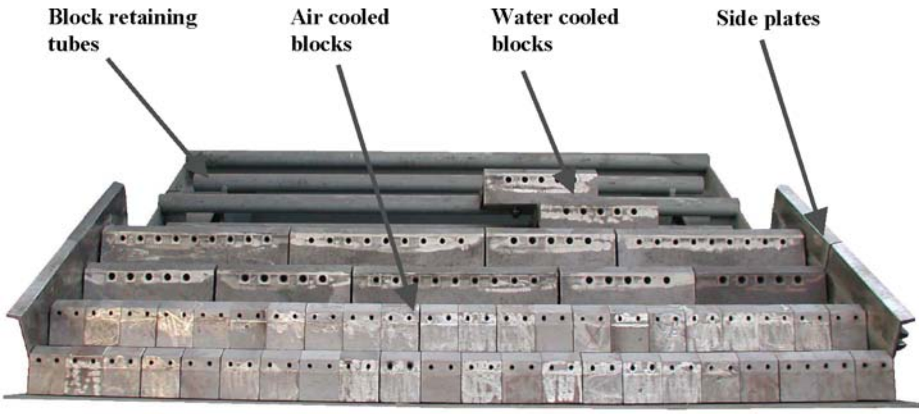
Hitachi Zosen Inova Technology. Figure 11
Grate module

The drive cylinder is attached to the grate carriage and secured to the support structure by means of a universal joint. The piston rod acts on the drive carriage attached to it, and thus to the movable grate block rows. The hydraulic cylinders are powered by a hydraulic station.

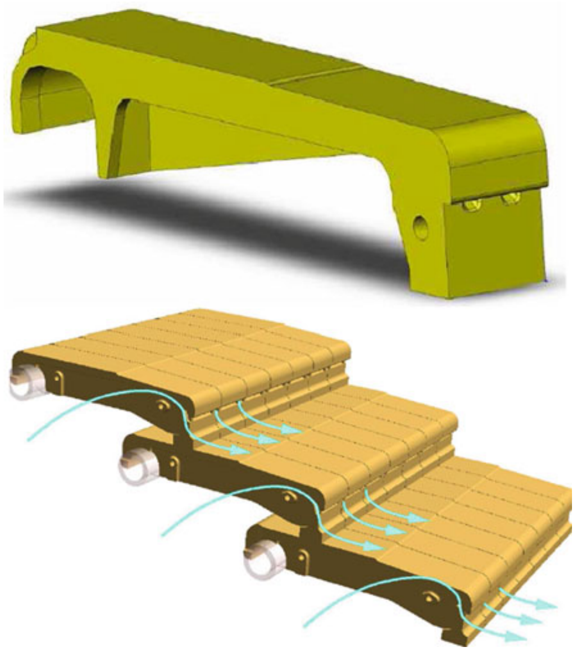
Grate Block Rows All of the six grate block rows per module are made up of either air-cooled or water-cooled grate blocks. The grate blocks are supported on the block-retaining tube as shown in Fig. 13. To minimize the gaps between blocks, they are held together by means of a tension rod. The two-part tension rod is inserted into the outermost grate blocks and tightened in the middle with a turnbuckle. This reduces riddlings and prevents undesirable airflow between the grate blocks and maintains a constant pressure drop across the grate.

Air-Cooled Grate Block The grate blocks support and push the waste through the combustion zone, and thus are exposed to extreme thermal, chemical, and mechanical stress. They are made of cast steel and are surface-machined as required.

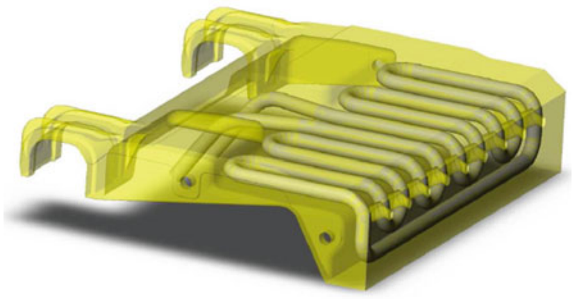
The grate blocks are also designed to distribute the primary air evenly through the waste bed. This is done



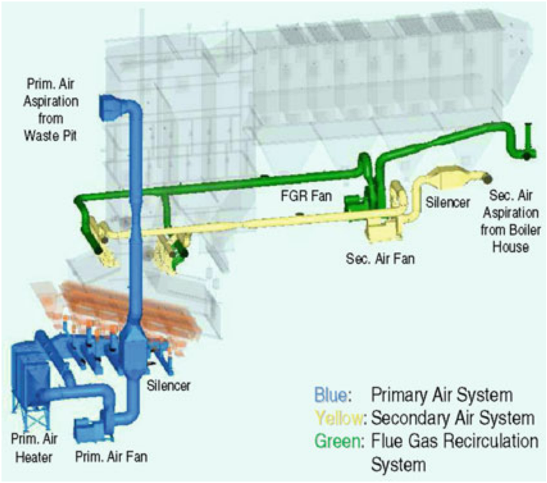
Hitachi Zosen Inova Technology. Figure 12
Rows of grate blocks in a module



Hitachi Zosen Inova Technology. Figure 13
Air-cooled grate block



Hitachi Zosen Inova Technology. Figure 14
Aquaroll® water-cooled grate block


Prim. Air Aspiration from Waste Pit
FGR Fan
Silencer
Sec. Air Aspiration from Boiler House
Sec. Air Fan
Silencer
Prim. Air Heater
Prim. Air Fan
Blue: Primary Air System
Yellow: Secondary Air System
Green: Flue Gas Recirculation System

Hitachi Zosen Inova Technology. Figure 15
Combustion air and flue gas recirculation systems

by openings at the front of the grate block. These openings are angled downward and protected by a small nose in order to help prevent them from getting fouled or clogged (Fig. 14).

To reduce the thermal stress on the grate block, it has to be cooled as uniformly as possible. This is achieved by the air that is guided through the grate block. The total opening of the air path through the grate blocks is designed such that a minimum pressure drop across the grate is maintained at all operation points. This assures an even distribution of the primary air through all grate blocks and reduces the risk of local superheating and associated excessive wear and tear. In addition, good distribution of primary air is essential to prevent excess formation of carbon monoxide and prevents the formation of clinkers on the grate surface.

Water-Cooled Grate Block For high-calorific wastes, the airflow alone does not provide sufficient cooling of the grate blocks. Aquaroll® water-cooled grate blocks as shown in Fig. 14 are used in such applications. Forced water circulation through a cooling tube cast into the grate block assures a low enough surface temperature of the grate block such that wear is drastically reduced.

Combustion Air System

General The main purpose of the primary air system (blue in Fig. 15) is to control and deliver the drying and combustion air to the hoppers underneath the grates at the right flow rate, pressure, and temperature as required for optimum combustion in the individual zones.

The secondary purpose of the primary air system is the ventilation of the tipping and waste pit area such that odors are contained therein.

The purpose of the secondary air system is to control and deliver the secondary combustion air to the combustion chamber above the grate. It creates the turbulence in the combustion chamber, which assures complete burnout of the flue gases.

Primary Air System

Overview The primary air is drawn in from the waste bunker and delivered separately, by the primary air fan, to the underside of the individual grate zones. The air is sucked from the pit in order to maintain a negative pressure and thus control odors. The system comprises the intake filter/screen, the primary air fan, the primary air preheater (if needed), as well as all associated ducts, dampers, and expansion joints. Distribution headers and ducts as shown in Fig. 16 direct the primary air to the respective grate modules and the combustion zones.

The primary air flow is controlled by the variable speed drive of the primary air fan. The distribution of primary air to each of the four zones is controlled by zone dampers. The flow rate of the primary air is measured in total as well as for each of the four zones to provide feedback to the combustion control system.

In areas where the calorific values of the waste are very low, the primary air is preheated using steam. This improves drying of the waste in the first grate zone and ensures proper burnout.

Hitachi Zosen Inova Technology. Figure 16
Primary air distribution headers and ducts underneath grate

Primary Air Fan The primary air fan is a standard centrifugal, backward inclined blade-type fan with single or double bearing and with direct connection to an electric motor drive. Since the air may contain dust from the waste pit area, the fan is equipped with openings for inspection and cleaning.

Primary Air Preheater For waste with low calorific value, a two- or three-stage air preheater is installed on the discharge side (positive pressure) of the primary air fan. The first stage preheats the air by cooling the condensate formed in the subsequent stage(s). The second stage is a low-pressure steam (LPS) and the third stage is a high-pressure steam (HPS) heater. The LPS stage is typically fed off an extraction from the steam turbine, while the HPS stage is feed off the boiler steam drum. This setup uses internal steam sources and therefore increases the plant's overall energy efficiency.

Since the primary air contains dust from the waste pit area, the open cross-sectional area is largely dimensioned, and the tube arrangements used as heat exchanger surfaces are exclusively in-line. A sufficient number of openings are provided for online inspection and cleaning. A bypass is provided for possible cleaning while the heater is off-line.

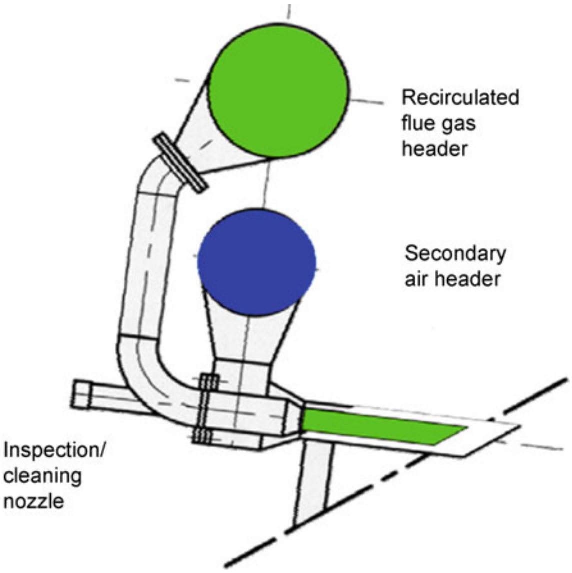
Secondary Air System

Overview Volatile components of the waste are released on exposure to heat. Some of them do not immediately combust on the grate because they may react more slowly or because oxygen is missing at that particular location in the waste bed on the grate. It is therefore necessary to provide overfire air and thoroughly mix the air with the flue gases above the grate to assure complete combustion of all volatiles released. This air is delivered via the secondary air system (Fig. 17).

The secondary air is drawn in from the boiler house under the roof in order to recover some energy from the radiant losses of the boiler equipment. It is delivered by the secondary air fan to the secondary air nozzles above the grate (Fig. 18). This system consists of the intake filter/screen, the secondary air fan, the secondary air preheater (if needed), the secondary air nozzles, and associated ducts, dampers, and expansion joints.

The secondary air is injected into the furnace at an angle, so that it causes a swirl flow in the combustion

Hitachi Zosen Inova Technology. Figure 17
Secondary air and recirculated flue gas headers and injection nozzles



Hitachi Zosen Inova Technology. Figure 18
Secondary air and recirculated flue gas nozzle

chamber. This swirl assures good mixing of the combustion gases and stabilizes the flue gas flow in the furnace.

The secondary airflow is controlled by the variable speed drive of the secondary air fan. Its flow is controlled by locally adjustable dampers in the distribution lines to the secondary air nozzles. These are preset during initial start-up. The flow rate of the secondary air is monitored and controlled by the CCS.

In case the calorific values of the waste are very low, the secondary air is preheated. This ensures that the combustion products are properly burned out without using auxiliary fuel.

Secondary Air Fan The same type of fan can be used as for the primary air system.

Secondary Air Preheater At very low waste heat values, a single-stage air preheater is installed on the discharge side (positive pressure) of the secondary air fan. It uses low-pressure steam. This setup extends the operating range of the furnace without using the auxiliary burners to maintain the furnace temperature. Using internal steam sources increases the plant's overall energy efficiency.

Flue Gas Recirculation System

General The purpose of the flue gas recirculation system (green in Fig. 18) is to supply a portion of the cleaned flue gases back to the combustion chamber, in order to improve mixing and provide additional cooling without the need for more combustion air. This improves the overall energy efficiency of the unit.

Description

Overview Flue gas recirculation reduces the amount of fresh air that has to be injected at the secondary air injection level for cooling and mixing purposes. It controls the combustion chamber temperature without adding air. This reduces the oxygen content in the flue gases as well as carbon monoxide and nitric oxides formation.

The flue gas is diverted downstream of the induced draft fan and is reinjected into the combustion chamber. The system comprises the flue gas recirculation fan as well as all associated ducts, dampers, and expansion joints. The system further has a purge connection for fresh air so the ducts can be purged with ambient air in order to dry them and prevent corrosion during a plant shutdown. The flue gas ducts must be well insulated.

The recirculated flue gases are injected through a central pipe of the secondary air nozzles. The flow rate is controlled by a variable speed drive as calculated by the CCS.

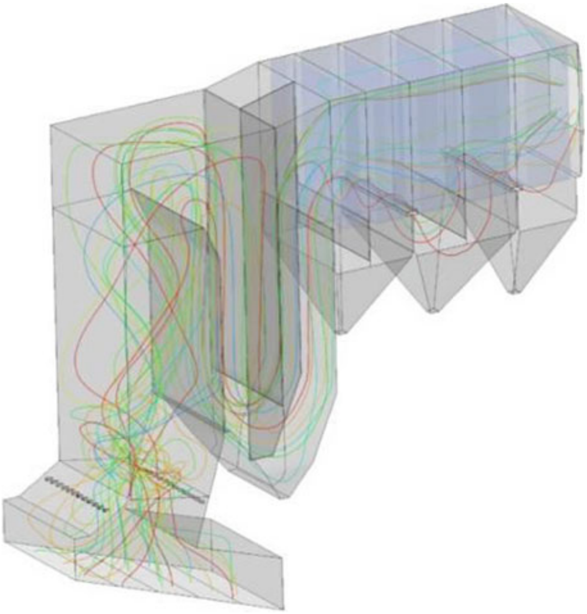
Recirculated Flue Gas Fan The same type of fan can be used as for the primary air system, however, adapted to the higher operating temperature and properly insulated.

Secondary Air/Recirculated Flue Gas Nozzles The nozzles (Fig. 18) are specially designed and placed to provide the impact needed to create a swirl flow pattern in the secondary combustion chamber. The concentric design with the recirculated flue gas in the center tube and clean secondary air in the annular gap minimizes the risk of plugs due to dust possibly present in the flue gases. A poke hole allows cleaning the center tube online.

Flow Optimization with Swirl Injection

General The injection of secondary air and recirculated flue gas as described above has been optimized to achieve best gas burnout. CFD modeling, as shown in Fig. 19, has been used to optimize number, location, and angle of injection points. Tangential injection into the furnace creates a swirl in the combustion chamber. This swirl improves the mixing of the combustion gases and helps maintain a uniform flow pattern.

Description The geometry of the secondary combustion chamber (SCC) is designed for optimal flow conditions. In addition, the arrangement of the secondary air and recirculated flue gas nozzles creates a swirl in the SCC. Due to this swirl, the gas flow is homogenized with respect to temperature, velocity, and concentrations (Fig. 20). This flow configuration results in:



Hitachi Zosen Inova Technology. Figure 19
CFD of swirl injection

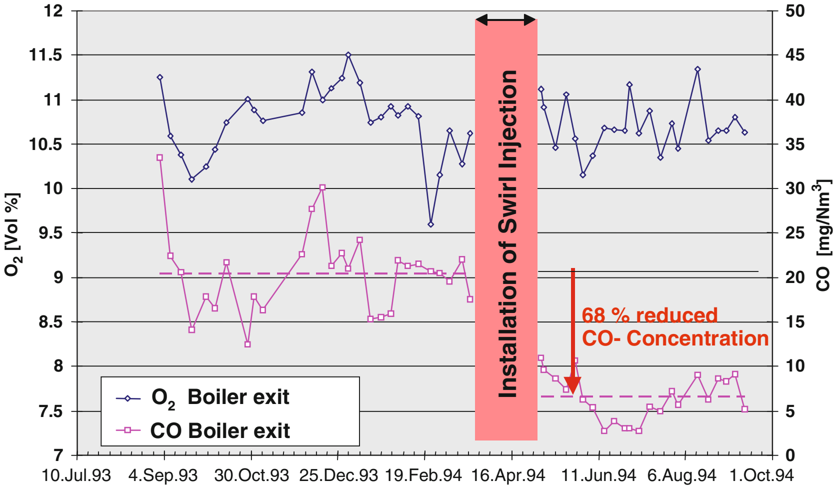
- Improved burnout of flue gas
- Uniform temperature profile across the secondary combustion chamber
- Reduction of CO-concentrations by approx. 70% (Fig. 20)
- Minimize risk of corrosion of unprotected heating surfaces
- Improved burnout of fly ash
- Reduced reformation of dioxins

Start-up Burners

General Start-up burners are specially designed for use in waste incinerators. They are located in the lower part of the combustion chamber above the secondary air injection area. Their primary purpose is to heat up the furnace gradually and in line with refractory requirements until the temperature is reached where wastes can be burned. The same applies during a controlled shutdown. During normal operation, the burners are switched off. Very seldom, with very low calorific wastes, the burners have to run to assure the minimum combustion temperature in the furnace.

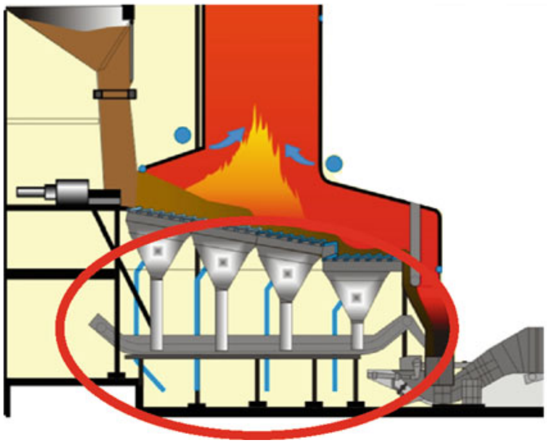
Description A separate burner fan provides the combustion air needed for the burner. The volume of combustion air is controlled by the burner control system and adjusted by the control damper incorporated into the air manifold. A pressure indicator is located in the duct upstream of the air manifold to monitor the operation of the combustion air fan. Since the burner is normally off during operation, it must be protected from the thermal radiation in the furnace. This can be done by a small, cooling air fan as shown on Fig. 21 together with the burner. Alternatively, a refractory-lined guillotine-type damper can be installed, which isolates the auxiliary burner from the furnace.

The burner control system ensures the burner safety and controls the correct fuel/air ratio. It is installed in a control cabinet located near the burner. All burner-operating commands, and the release signal from the system safety network, come from the plant control system. The local control panel can be used to



Hitachi Zosen Inova Technology. Figure 20
Average CO levels before and after installation of the swirl injection system

Hitachi Zosen Inova Technology. Figure 21
Typical auxiliary burner with cooling air fan



Hitachi Zosen Inova Technology. Figure 22
Riddlings conveyor

commission and optimize burner operation as well as to cancel local error messages during operation.

Conveyor of Grate Siftings/Riddlings

General While all moving parts of the ram feeder and the grate are designed with minimum gaps, the thermal expansion of the materials between ambient and operating temperatures unavoidably lead to small gaps through which small parts in the waste or the ash can fall through. The purpose of the grate siftings/riddlings conveyor (circled in Fig. 22) is to constantly remove these so-called siftings/riddlings while maintaining the air seal to the environment for the primary air system.

Description A wet drag chain conveyor is used for this application. One conveyor can be used for up to two grate lanes. Riddlings are collected from five locations for each lane: four grate modules and the ram feeder. The siftings/riddlings are collected in hoppers underneath these locations and delivered to the conveyor via chutes. The chutes are immersed under the water level inside the conveyor, thus ensuring an air seal between the furnace enclosure and the environment. The wet drag chain conveyor quenches the grate siftings and transports them toward the bottom ash extractor.

The amount of siftings/riddlings delivered varies and can change over the operating years due to wear and tear of the moving equipment. The drag chain conveyor can be adjusted to these conditions. Appropriate openings are installed at various points of the conveyor for inspection and cleaning as needed. The conveyor is a watertight steel construction. The key areas prone to wear are protected using wear-resistant plates.

The conveyor designed for the case that a heavy piece such as a damaged grate block falls into the conveyor. A shear pin protects the conveyor, and a zero-speed switch alarms the control room. The conveyor's water level is maintained using a float valve.

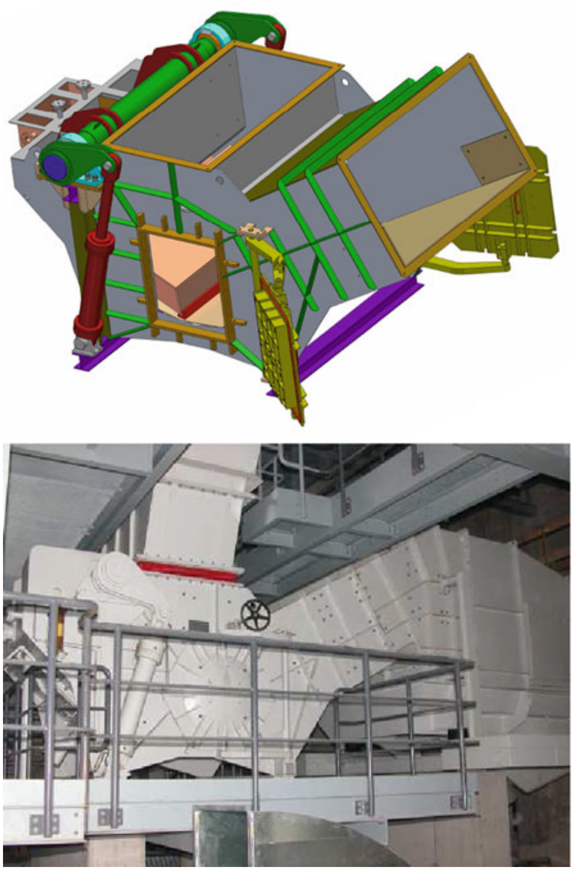
Bottom Ash Extractor

General The purpose of the bottom ash extractor (Fig. 23) is to cool the bottom ashes from the furnace as well as discharge bottom ashes and riddlings while at the same time maintaining the air seal from the furnace to the environment.

Description The bottom ashes remaining at the end of the grate fall into the bottom ash extractor, where they are quenched in a water bath. The grate siftings/riddlings are delivered to the bottom ash extractor via the grate siftings/riddlings conveyor. The bottom ash discharge chute has an airtight connection to the grate, and the water bath in the bottom ash extractor provides an air seal between the furnace and the environment.

The design of the bottom ash chutes and extractor is such that the large pieces of bottom ash can be discharged without difficulty. A hydraulically driven ram pushes the bottom ashes once every few minutes from the bottom of the extractor upward to the sloped chute and toward the outlet. There, the ashes stay until they are further pushed by the next ram stroke. After a few strokes, the ashes reach the edge at the outlet, fall on a belt or vibrating conveyor, and are transported to the ash storage and treatment. During this period, water contained in the ashes drains back to the water bath so that the final moisture content of the bottom ash is typically about 20%.

The hot ashes falling into the water bath cause some of the water to evaporate. This water vapor rises into the combustion chamber through the bottom ash chute. Water vapor is also formed in the outlet chute of the ash expeller since larger ash pieces may still be warm. From there, the vapor is conveyed into the furnace by means of a small fan. A water-level control regulates the feed of water to replace the evaporated water and maintain the water level in the bottom ash extractor.



Hitachi Zosen Inova Technology. Figure 23
3D model and picture of bottom ash extractor

The bottom ash extractor is made of reinforced steel plates in a completely welded design. In order to reduce wear, the trough is internally lined with wear plates. The bottom ash discharge chute is a double-walled structure made of steel plate.

Hydraulic Station

General The purpose of the hydraulic station is to power the hydraulic cylinders for the feed hopper, the ram feeder, and the grate and the bottom ash extractor.

Description A single hydraulic station as shown in Fig. 24 is provided for each incineration train. The pumps, tanks, heat exchanger, and valves are all mounted and placed in a secondary containment in an easily accessible location at grade. Hydraulic oil is supplied by a set of hydraulic pumps located on the hydraulic reservoir. Each hydraulic drive has a separate hydraulic control block with electrical control elements.

Two redundant sets are equipped with a pressure-dependent flow rate control system, to perform the respective functions of the feed hopper, the ram feeder and the grate. Each grate module has a separate control block.

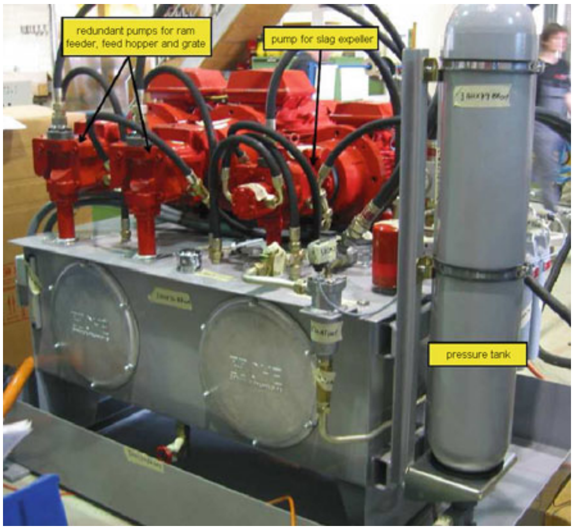
A third hydraulic pump is pressure-controlled and is used for the bottom ash extractor. Redundancy is provided by the other two pumps.

A water-cooled heat exchanger is integrated into the return line of the hydraulic circuit.

A pressurized tank is provided for the case of power failure in order to safely close and/or open the necessary cylinders and bring the system into its fail-safe position.

Combustion Control System

General The purpose of the combustion control system (CCS) is to provide a stable steam load in automatic mode. Since waste is a highly variable fuel, this is not a simple task. Various parameters must be monitored and various equipment need to be controlled in a concerted way.



Hitachi Zosen Inova Technology. Figure 24
Typical hydraulic station

Description

Overview Control Functions Figure 25 shows a typical Hitachi Zosen Inova combustion control system (CCS). It allows for largely automatic and secure operation at the requested load. Even at varying waste qualities, compliance with legal operating conditions such as the combustion chamber temperature and the oxygen content in the flue gas as well as an efficient gas and bottom ash burnout are assured because of the automatic control system. Steady control of the combustion conditions minimizes stress to the equipment and reduces fouling in the combustion chamber. Good combustion control includes:

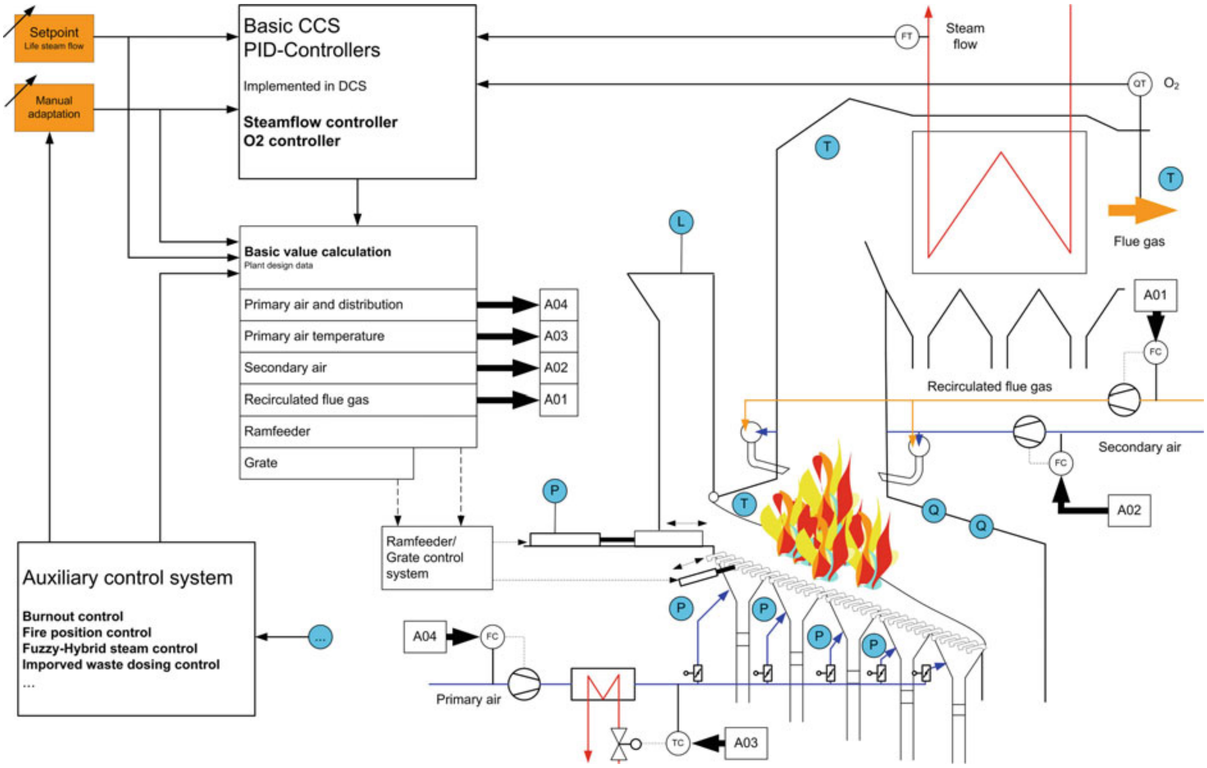
- Keeping the steam flow constant, as preset by the operator
- Keeping the flue gas flow rate constant
- Ensuring an efficient gas burnout by keeping the oxygen content in the flue gas within the desired limits
- Ensuring the fire is in the right position for an efficient bottom ash burnout

All these aspects are interdependent and are heavily influenced by changing waste properties. Thus, a multivariable control system is required. The following parameters are monitored and controlled:

- Steam flow rate
- Oxygen content at the boiler outlet
- Primary airflow rate
- Primary air temperature
- Secondary airflow rate
- Flue gas recirculation flow rate
- Temperature of ceiling above grate end

Also, the following equipment is controlled:

- Ram feeder
- Grate
- Primary air fan
- Primary air dampers per zone
- Primary air heater
- Secondary air fan
- Flue gas recirculation fan

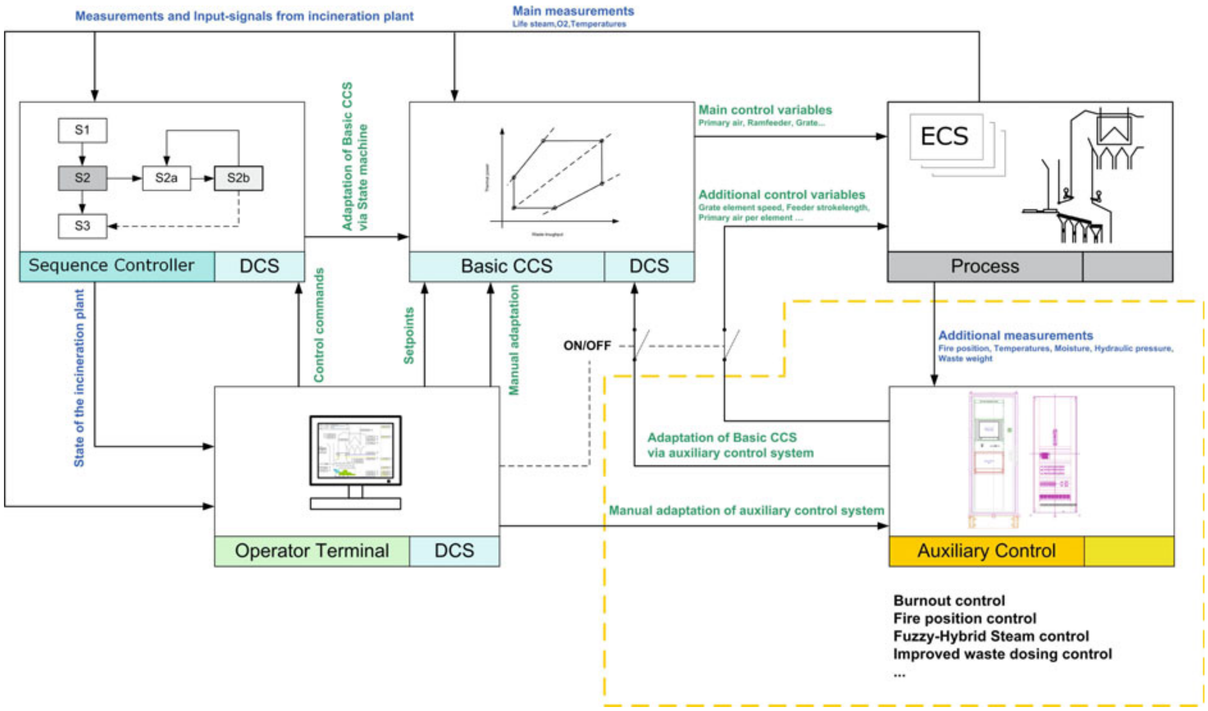


Hitachi Zosen Inova Technology. Figure 25
Overview main combustion control functions

The control structure consists of both parallel connections of different controllers and of serial (cascading) connections of controllers located at various levels in the control system.

Overview Control Systems As shown in Fig. 26, the DCS controls the EfW plant. It contains three main elements: (1) the basic Combustion Control System (CCS) as well as other control systems of the plant such as the controls of the Air Pollution Control System (not shown in the graph). (2) the Sequence Controller, which informs about the state of the plant as well as automatically brings it to the next mode as needed. (3) the Operator Terminals where the plant can be monitored and manually optimized or operated. The DCS receives signals from the plant such equipment status and flowrates or temperatures and it controls the plant through main control variables such as primary air temperatures and flowrates or ram feeder throughput etc. Its parameters and setpoints may not only be determined through the Operator Terminal, but also by Auxiliary Control Systems (ACS) which further improve quality and stability of operation.

The main control variables either directly control certain parameters such as for example the steam flow to the primary air heater such that the correct primary air temperature is reached, or it controls Equipment Control Systems (ECS) which are in turn controlling various individual equipment functions to assure the desired operation. The DCS will for example ask the ECS of the grate and ram feeder to provide more or less waste to keep the steam flow to the turbine constant. But it lets the ECS do the detailed equipment control to feed, move and riddle the waste fuel by setting the stroke of all grate elements in the individual combustion zones. The ideal stroke pattern can be assigned to each individual grate element depending on waste properties and actual combustion profile. The optimum stroke frequency is preset for each grate module. For example, the grate modules in the main combustion zone have a higher stroke rate than those in the


Measurements and Input-signals from incineration plant
Main measurements
Life steam,O2,Temperatures
S1
S2
S2a
S2b
S3
Sequence Controller
DCS
Adaptation of Basic CCS via State machine
Basic CCS
DCS
Main control variables
Primary air, Ramfeeder, Grate...
Additional control variables
Grate element speed, Feeder strokelength, Primary air per element ....
ECS
Process
State of the incineration plant
Control commands
Setpoints
Manual adaptation
ON/OFF
Additional measurements
Fire position, Temperatures, Moisture, Hydraulic pressure, Waste weight
Adaptation of Basic CCS via auxiliary control system
Manual adaptation of auxiliary control system
Operator Terminal
DCS
Auxiliary Control
Burnout control
Fire position control
Fuzzy-Hybrid Steam control
Improved waste dosing control
...

Hitachi Zosen Inova Technology. Figure 26
Main combustion control system architecture

burnout zone. For simple changes of the throughput the stroke rates for the entire grate can be uniformly increased or decreased without changing the relationship between the individual grate modules. If there is a change in calorific value and therefore in the combustion profile on the grate, the specific stroke rate for the individual grate zones can be adjusted by changing the frequency or length of strokes. This affects the travelling speed of the waste on the grate, but also the thickness of the waste layer on the grate, the stoking effect and therefore the combustion profile.

For an even more precise and automatic control of the combustion profile, specially for large grates and very inhomogeneous wastes one or more ACS can be installed. The Burnout Control or the Fire Position Control measure the location of the fire by means of pyro-sensors at various locations across the width and along the length of the grate. This allows to automatically adjust the stroke pattern locally on the grate if the waste properties are different on one side of the grate than on the other side. This further improves the burnout of the waste to even lower values. With the basic CCS one typically can only see that the waste properties have changed after the steam flow has already dropped or increased. An improved waste dosing control and fuzzy-hybrid steam control allow to better anticipate possible changes in the waste composition and to proactively change the control settings such that the steam flow to the turbine and therefore the power production constantly remains at the desired value.

Basic Control Functions The basic settings for the controlled equipment are set during design and adjusted during commissioning. These settings are valid for the desired operation point with mean waste properties. During normal operation, there may be two reasons why these settings have to be changed:

The operator needs to operate at a different operation point because he receives more or less waste or because he wants to deliver more or less electricity or steam.

The operator wants to stay at the same operation point, but the properties of the wastes change with time, for example, because the ratio between industrial and residential wastes changes or because the waste is wetter due to heavy rain in the past days or similar.

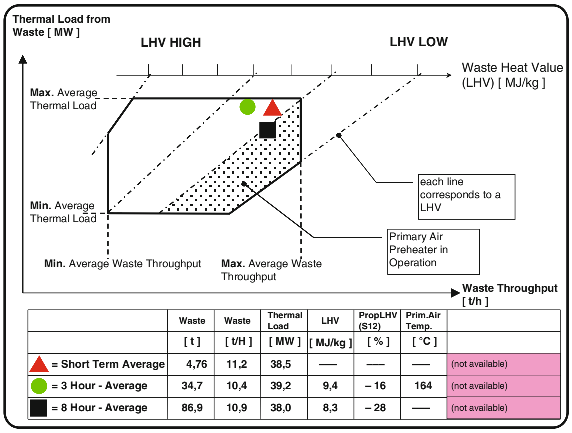
The main parameter to be controlled is the produced steam flow, while keeping other parameters such as the temperatures and the oxygen content in the flue gases within the desired range. When the steam flow is below the desired value at a given load, the primary airflow below the grate is increased; simultaneously, the secondary airflow above the grate is decreased by the same rate. With this intervention, the combustion load is increased at constant total air supply. At the same time, the ram feeder and the grate stroke frequency are increased. As a result, the fire is intensified and more waste is fed to the grate. If the capacity is increased or if the oxygen concentration in the flue gas drops, the total combustion airflow is increased and the recirculated flue gas flow adjusted. In analogy, when the steam flow is higher than desired, the control interventions are reversed.

Hitachi Zosen Inova Navigator The CCS automates the complete combustion process in such a way that the operator has to select only the set point for the desired steam flow rate and an estimation of the calorific value of the present waste (calorific value adjustment). In case of a change of the waste qualities, all basic set points are simultaneously modified by means of the calorific value adjustment so that favorable combustion conditions are always given.

Using measured data and some minor assumptions, the thermal load from waste in the combustion diagram can be calculated (Fig. 27). The measured waste flow, at the crane, is used to calculate the actual calorific value and set the next operating point (short term, 3 h and 8 h average). The navigator screen is permanently updated.

While this screen allows the operator to see the past operating points, there is also a feature called "calorific value adjustment" in the CCS that allows the operator with one simple adjustment to adjust the system for an anticipated change of the future operating point. The CCS then correctly adjusts the many set points of all the relevant subsystems.

Additionally, manual intervention is possible allowing the operator to further optimize the combustion process if necessary. However, the Hitachi Zosen Inova CCS is designed so that the plant cannot be continuously operated outside the permissible range as defined by the combustion diagram.


Thermal Load from Waste [MW]
LHV HIGH
LHV LOW
Waste Heat Value (LHV) [MJ/kg]
Max. Average Thermal Load
Min. Average Thermal Load
each line corresponds to a LHV
Primary Air Preheater in Operation
Min. Average Waste Throughput
Max. Average Waste Throughput
Waste Throughput [t/h]

	Waste	Waste	Thermal Load	LHV	PropLHV (S12)	Prim.Air Temp.	
	[t]	[t/H]	[MW]	[MJ/kg]	[%]	[°C]	
▲ = Short Term Average	4,76	11,2	38,5	—	—	—	(not available)
● = 3 Hour - Average	34,7	10,4	39,2	9,4	– 16	164	(not available)
■ = 8 Hour - Average	86,9	10,9	38,0	8,3	– 28	—	(not available)

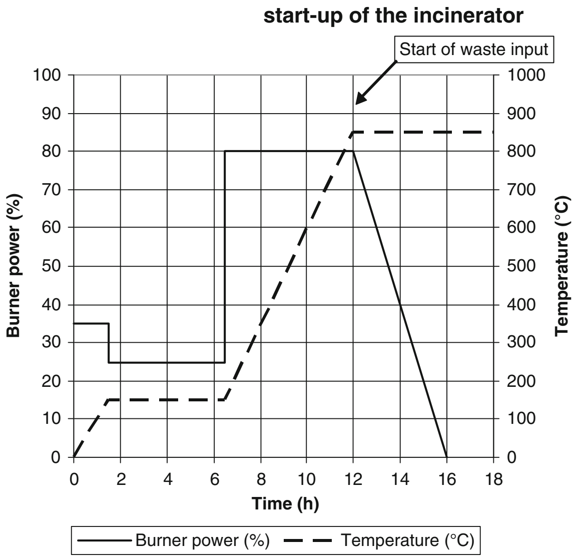
Hitachi Zosen Inova Technology. Figure 27
Navigator diagram

Start-up Sequence Also start-up is automated in the CCS. This includes heating up the post-combustion chamber by means of the start-up burners following a given ramp (Fig. 28) until the temperature in the chamber reaches the required minimal temperature. Once this condition is met, the gate in the feed chute is opened to admit the waste onto the grate where it will start to burn immediately.

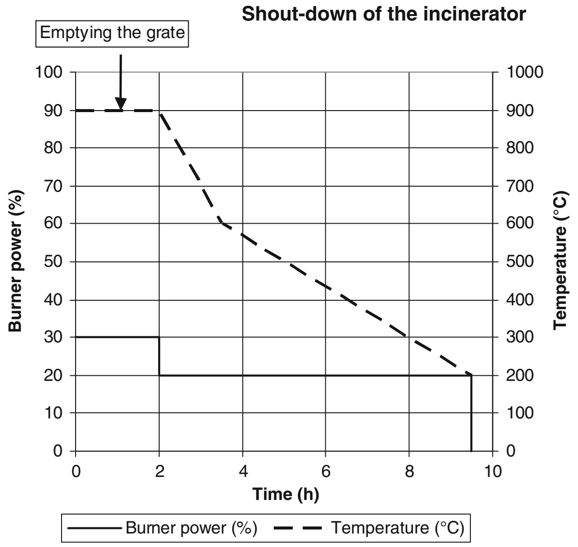
Shutdown Sequence After the waste feed is stopped, the temperature of the post-combustion chamber is kept above the minimal required temperature using the burners. When the grate is empty, the temperature in the post-combustion chamber is gradually and steadily decreased as shown in Fig. 29, until it is safe to start maintenance work.

Steam Boiler

General The purpose of the steam boiler is to transfer the heat from the flue gas to water and steam. The most frequent case is to produce superheated steam, as required in a steam turbine / generator to produce electricity. In the case where the recovered energy is used for heating only, a boiler can be designed to produce saturated steam only or heat water to be used in district heating.



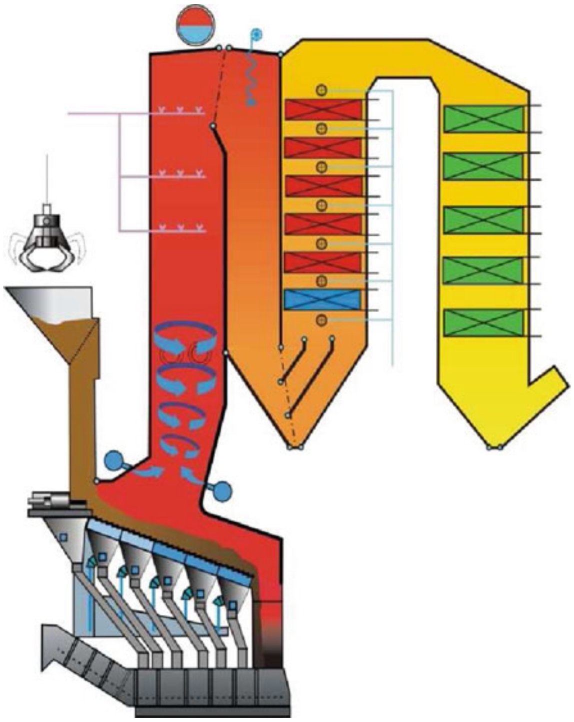
Hitachi Zosen Inova Technology. Figure 28
Typical start-up curve



Hitachi Zosen Inova Technology. Figure 29
Typical shutdown curve

Description The Hitachi Zosen Inova steam boiler, in the example of the four-pass vertical boiler shown in Fig. 30, can be described by the following main functions it provides in the water–steam cycle:

- *Economizer* (*green bundles*): In this section, the pressurized feed water is preheated to close to boiling conditions and fed into the steam drum.



Hitachi Zosen Inova Technology. Figure 30
Typical four-pass steam boiler

- *Evaporator* (*blue bundles and all boiler walls*): In this section, water is recirculated from and to the steam drum. It is heated in the evaporator surfaces such that it partially evaporates. Steam is separated in the steam drum and water is recirculated again.
- *Superheater* (*red bundles*): In this section, the steam separated in the steam drum is further heated to desired conditions.

The steam boiler can also be described by its configuration in terms of the number and arrangement of passes. A pass is a section of the boiler in which the flue gases flow in the same direction. The boiler can be designed in various possible configurations. In the following, a four-pass vertical boiler is described in further detail. It consists of the following passes in which one or more of the above functions are performed:

- *First pass:* Vertical radiation pass made of evaporator membrane walls with refractory lining and/or Inconel cladding on flue gas side.

- *Second pass:* Vertical radiation pass made of evaporator membrane walls and possibly with additional evaporation bundle sections.
- *Third pass:* Vertical convection pass made of evaporator membrane walls as well as with bundles of superheater and evaporator sections.
- *Fourth pass:* Vertical convection pass with steel casing and bundles of economizer sections.

The steam boiler (passes 1–3) can be bottom or mid supported or even suspended. The fourth boiler pass is bottom supported. The steam drum is placed above the first boiler pass on the steel structure.

The first pass is located above the incinerator grate and consists of membrane walls. Since these are exposed to the hottest temperatures and to the fire, they have to be protected by refractory lining or inconel clading. The sidewalls of the boiler extend downward such that the walls on the side of the grate form an integral part of the first pass. The front and back walls of the first pass extend downward to create the sloped roofs above the grate. The differences of thermal expansion between the boiler and the grate as well as between the third and fourth pass are compensated by expansion joints.

Auxiliary Boiler Systems The steam drum is the central part of the steam boiler, where water and steam is separated and the water level and the steam pressure are monitored. It includes a feed water preheater coil as well as separating baffles or cyclones to efficiently separate water from the steam. It is of fusion-welded construction and fabricated from steel plate. It contains a manhole, hinged to open inward, as well as all necessary nozzles, connections, and openings required for operation, maintenance, and testing. All drum nozzles for instruments, valves, fittings, and tubes are welded to the drum.

Since the boiler load may vary and the boiler heating surfaces are subject to scaling, the uncontrolled superheater section may provide more or less superheating depending on load and conditions. To be able to control the superheated steam temperature at all conditions, one or two spray desuperheaters are installed between the various superheater sections. There, feed water is injected in order to accurately adjust the superheater steam temperature.

The superheated steam is collected in an outlet header where a pressure control valve and a stop valve are installed before the live steam pipe is directed toward the steam turbine.

Safety valves are provided on the steam drum and in the live steam pipe to protect the boiler from excessive pressures.

The cleaning of surfaces causes the ash to fall down from refractory, membrane walls or tube bundles. The ash is collected in the ash hoppers below each boiler pass and then removed via double gate or rotary valve to the ash collection system.

All necessary provisions for normal maintenance access are made. Access doors are placed where required on both sides of the boiler. Ample space is provided between each tube bundle for adequate inspection and cleaning. External platforms provide access to each door and all control instruments. Provisions are made so internal platforms can be installed at various boiler levels for inspection, cleaning, and maintenance during outages.

Boiler Cleaning

General A waste incineration boiler is inevitably exposed to high fly ash loads. Hence, the effectiveness of the heat exchanger surface is reduced during the operation time by fouling. The purpose of online cleaning of the heat exchange surfaces is to extend the operating period and reduce maintenance cost. This can be accomplished by various means. The main options are:

- Water shower cleaning
- Sootblowers
- Rappers

Description

Water Shower Cleaning A water shower cleaning system as shown in Fig. 31 can be installed for the online cleaning of the empty boiler pass. It enables an automatic cleaning of the membrane walls by means of water injection. The evaporation of the water on the dust layers results in sudden cooling and contraction of the ash layer which lets the ash layer crack and fall off.

The water shower system is not often used in the beginning of an operation campaign. During that phase, the clean surfaces of the evaporator sections in pass 1 and 2 take more heat out of the flue gases than in the average. So less heat is available in the superheaters in the third pass and they may not reach the required superheated steam temperature at all conditions.

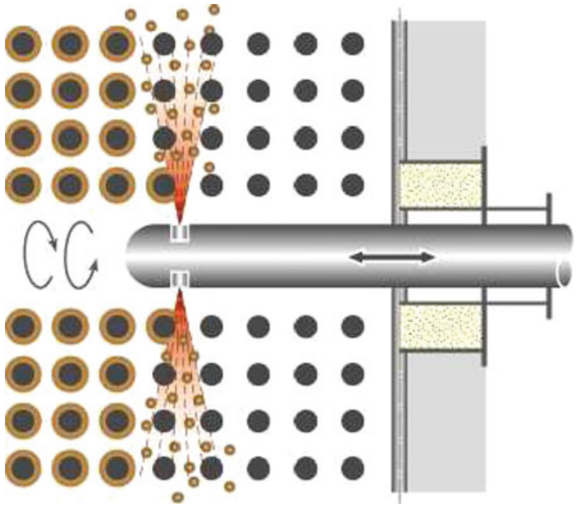
With proceeding of the operation campaign, it is used more and more often to assure that the flue gas temperatures at the inlet to the superheaters do not exceed design temperatures. This is usually about once per week at the end of the operating campaign. And it is used right before shutdown for manual boiler cleaning which makes the cleaning much easier and faster.

Sootblowers Sootblowers as shown in Fig. 32 are installed in before, after, and between bundles in a vertical boiler pass.

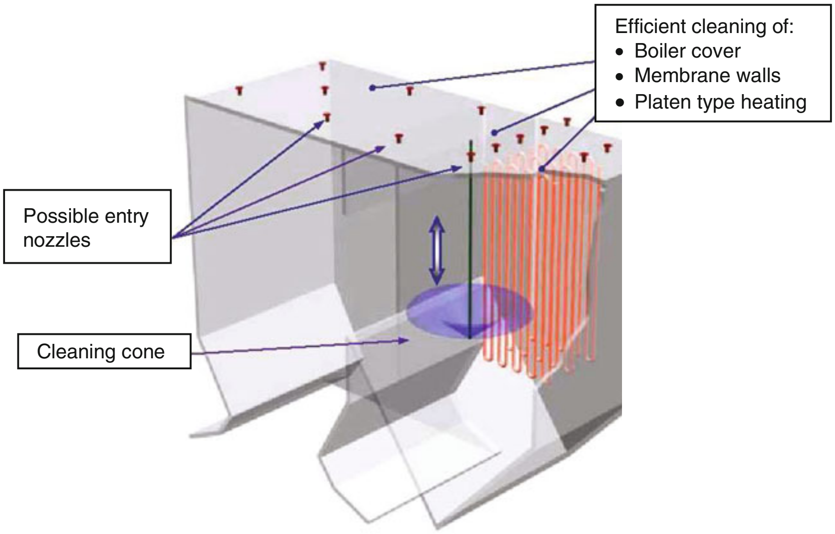
The areas of the tube bundles exposed to the direct sootblower spray are protected by stainless steel tube shields. Various types of sootblowers exist. A typical retractable-type sootblower has two opposite high-performance nozzles and is helically moved in and out of the boiler during a cleaning cycle. Between cycles it is located in a cavity of the boiler wall and cooled with a small air stream to avoid prolonged exposure to the high temperatures.

Sootblowers allow an effective online cleaning of tenacious fouling at heating surface banks with high flue gas temperatures. Sootblowers use superheated steam or compressed hot air as blowing agent because wet and moist steam can cause further caking of the fouling and cold air can cause corrosion if surface temperature of the sprayed area falls locally below the acid dewpoint of the flue gases.

Rappers Rappers are usually used in vertical tube banks installed in a horizontal boiler pass. One rod connected to each of the vertical tube rows penetrates the boiler wall. On the outside of the boiler this rod can



Hitachi Zosen Inova Technology. Figure 32 Retractable sootblower



Hitachi Zosen Inova Technology. Figure 31 Typical water shower system be hit with a hammer from time to time. The rod transmits the impact of the hammer to the heat exchanger tubes and creates a vibration such that the accumulated dust and scaling falls off. The key is to hit the bundle at the right place with the right force to achieve ideal vibrations and acceleration.

The hammering of the many heat exchanger tube rows is automated. Figure 33 shows a traveling rapping system which moves along the boiler and strikes each rod at a given interval and with a defined force.

Boiler Surface Protection

General The purpose is to protect the boiler surfaces from flame impingement and aggressive flue gas and boiler ash conditions. Usually, a combination of refractory lining and Inconel cladding is used.

Refractory Lining The refractory is different in various sections of the furnace (Fig. 34) and is optimized for:

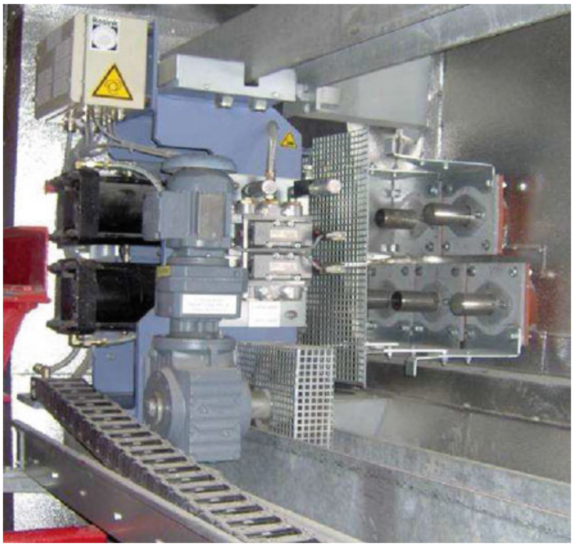
- Proper ignition of waste
- Improved burnout of waste on the grate
- Protection of heating surfaces from furnace erosion
- Protect heating surfaces from high flue gas temperatures (corrosion protection)
- Achievement of high residence time of the flue gases at high temperatures for a complete burnout

These requirements are satisfied by choosing appropriate materials and thickness of refractory in different zones.

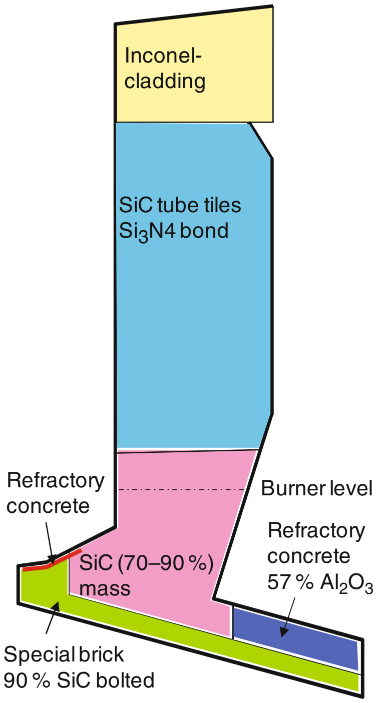
Inconel Cladding Inconel cladding is one option to protect the steel surfaces of the boiler from the corrosive atmosphere of the flue gases after waste combustion. Its biggest advantage is that it protects the boiler tubes without reducing the heat transfer. Its main disadvantage is cost. This is why the design of the surface protection concept is always an optimization which has to be done in view of the waste properties and the plant owner's operation and maintenance philosophy. A welding robot applying an Inconel overlay to a boiler membrane wall is shown in Fig. 35.

Sampling Station

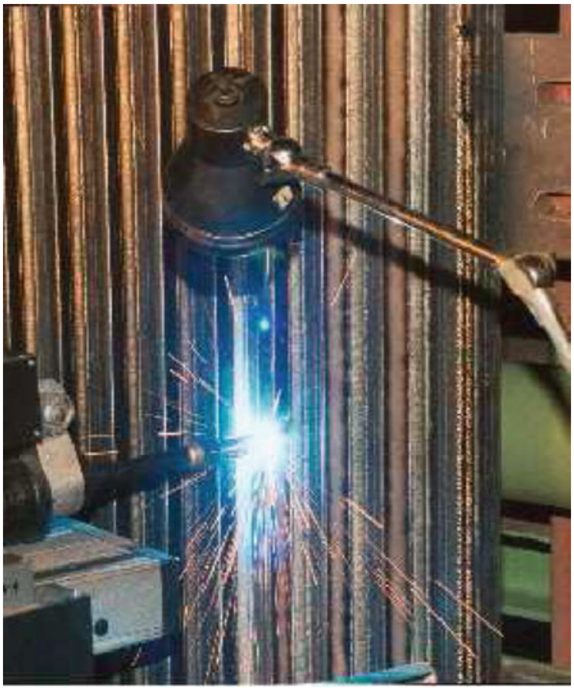
General The purpose of the sampling station is to monitor the water and steam quality at various locations in the boiler system (Fig. 36). Correct sampling



Hitachi Zosen Inova Technology. Figure 33
Automated rapper system



Hitachi Zosen Inova Technology. Figure 34
Surface protection concept in furnace and first boiler pass



Hitachi Zosen Inova Technology. Figure 35
Welding roboter applying Inconel overlay to boiler tubes

and possible required adjustments prevent corrosion or other serious malfunction of the water side of the boiler or subsequent equipment.

Description The online sampling of a boiler is typically made to analyze the following fluids and parameters:

- Condensate (pH and conductivity)
- Boiler water (pH and conductivity)
- Steam (conductivity)
- Feed water (oxygen, pH, and conductivity)

To simplify analysis, all fluids are piped to an easily accessible location. Therefore, sampling station equipment contains a lot of piping, valves, and instruments, as well as heat exchangers to cool the fluids to temperatures suitable for handling.

Chemical Dosing

General The purpose of the chemical dosing system is to condition the water–steam cycle by the addition of chemicals such that required water and steam properties are maintained.

Hitachi Zosen Inova Technology. Figure 36
Sampling station

Description The water–steam cycle chemistry is controlled in accordance with the requirements in the relevant boiler codes and turbine manufacturers by chemical injection into the feed water, usually of ammonia solution and trisodium phosphate.

Ammonia is fed continuously into the feed water pipe between the feed water tank and the feed water pumps. Trisodium phosphate is fed continuously into the feed water pipe before the economizer of the boiler, but after the connecting pipes for spray water for the cooling of desuperheaters.

Turbine System

General The purpose of the turbine and generator system is to take the superheated steam produced by the boiler and create electrical energy.

Description The turbine system is a standard steam generator turbine package consisting of an extraction turbine, a generator with electrical equipment and

Hitachi Zosen Inova Technology. Figure 37 Typical steam turbine and generator

protection, a lubrication oil system including oil filter and cooler and local control panel (Fig. 37).

The turbine will typically have one or two steam extraction points. The first is typically around 3–5 bars (about 45–75 psi), which will provide steam to various internal, and possibly to external steam consumers via the low-pressure steam (LPS) system. A second extraction point is made if heat is used by a heat consumer at higher temperature. This extraction point is typically around 10–25 bars (150–375 psi), which will provide steam to external and possibly to some internal steam consumers via the medium pressure steam (MPS) system. If there is no second extraction point, any possible internal MPS consumers will be fed with high-pressure steam directly from the boiler. The exhaust steam from the turbine is directed to the condenser at a pressure as low as possible with prevailing ambient conditions.

During normal plant operation, the pressure control in the live-steam header will be performed by the steam turbine's pressure control valve. In case of a trip of the steam turbine, a controlled live-steam release to the condenser via the turbine bypass system will be initiated, in order to prevent the boilers' safety valves from relieving pressure. In the event that the steam turbine's capacity (typically 110% of the nominal load) is exceeded, the turbine bypass system is also activated, which will then take over pressure control in the main steam header. The turbine bypass system is also active during start-up and shutdown of the boiler in order to minimize any water losses. If the controlled live-steam release or the turbine bypass does not work, steam is blown off to the air through safety valves.

Condenser

General The purpose of the condenser is to condense the turbine's exhaust steam during normal operation. However, during start-up, shutdown, and/or upset conditions, the purpose of the condenser is to condensate the entire steam load.

Condenser Types Condensers can typically be one of the following three types:

- Non-evaporative water-cooled condenser
- Evaporative water-cooled condenser
- Air-cooled condenser

The lower the condensation temperature which can be achieved, the higher is the electrical energy efficiency of the turbine and the EfW plant.

A once-through non-evaporative water-cooled condenser will reach the highest efficiency. However, it is rare today to find a large enough body of water which can absorb the cooling needs of a power plant seldom today there is a large enough water carrier close to the EfW plant which is allowed to be heated up for cooling of a power plant. This is why this solution cannot be realized very often.

An evaporative water-cooled condenser reaches average efficiencies of the three proposed solutions. Its downside is the requirement of treated water which is not available everywhere in sufficient amounts and the frequent formation of a vapor plume which is not welcome everywhere.

So an air-cooled condenser as shown in Fig. 38 remains as the least energy-efficient option which however also causes negligible impacts to the environment and the neighborhood.

Description of an Air-Cooled Condenser Large air fans blow ambient air across the steam condenser tubes to cool them.

The condenser is designed both for the maximum steam flow from the turbine in normal operation as well as for the total plant steam flow in the turbine bypass case.

Hitachi Zosen Inova Technology. Figure 38
Air-cooled condenser

Air or other incondensable gases entering the condenser with the steam are extracted from the steam space of the condenser by a two-stage steam jet ejector. The condensed steam is returned to the condensate tank. The air extracted by the ejector is vented to the atmosphere. The driving steam for the ejector is taken from the HPS header.

Cooling System

General The purpose of the cooling system is to adequately cool the turbine and generator's lubrication system.

Description The cooling system is a closed loop cooling system that provides cooling water to the turbine oil cooler, the generator air cooler, the compressor cooling, and the sampling coolers. In an air cooler, the excess heat from the cooling water is transferred to the ambient air supplied by fans. The air cooler consists of different small cooler modules. Each cooler module has a fan with electrical motor.

In case of partial-load operation and/or low ambient air temperatures, the cooling-water outlet temperature can be adjusted by taking some of the cooler modules out of operation. The cooling water is circulated by means of redundant cooling-water circulating pumps. The cooling-water system is filled and/or drained by the cooling-water system filling pump. It is important to keep the pressure constant in the system and this is done by means of a diaphragm pressure-compensating tank equipped with a nitrogen buffer.

Hitachi Zosen Inova Technology. Figure 39
Deaerator and feed water tank

Feed Water System

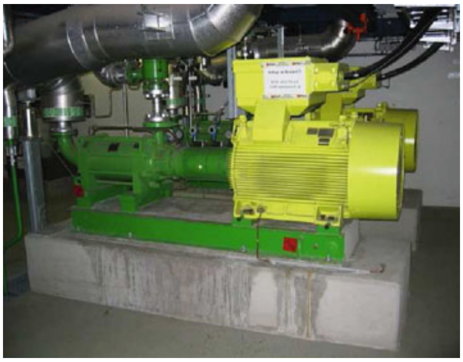
General The purpose of the feed water system is to provide the boiler and superheaters with the required feed water flow and pressure.

Description The feed water system consists of the deaerator with feed water tank (Fig. 39), the redundant feed water pumps (Fig. 40), and all the associated piping and valves. The feed water tank is sized adequately to provide enough storage volume to provide feed water to the boiler for a defined time.

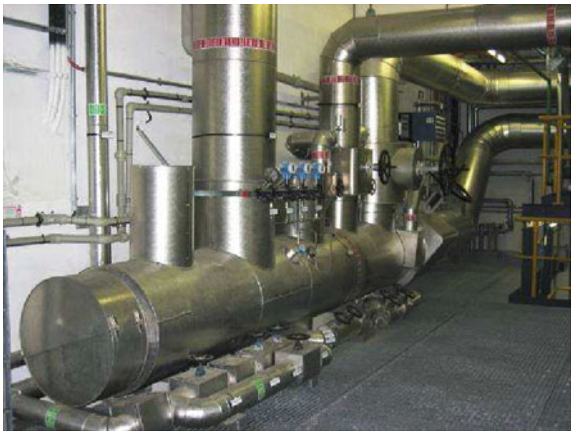
High-Pressure Steam System

General The purpose of the high-pressure steam (HPS) system is to deliver HPS to the in- and out-of-plant consumers as required.

Description The boiler produces high-pressure superheated steam which is sent to the turbine/generator to create electrical energy during normal operation. In order to compensate for possible variations in load, the steam is first sent to the HPS system. The HPS system typically consists of the live steam piping from the boiler to the HPS header, the HPS header



Hitachi Zosen Inova Technology. Figure 40
Feed water pump



Hitachi Zosen Inova Technology. Figure 41
HP steam header

(Fig. 41), the piping from the HPS header to the turbine, the piping from the HPS header to the air-cooled condenser via steam pressure reduction station of the turbine bypass, the piping from the HPS header to the LPS header via the steam pressure reduction station (HPS/LPS), and the piping from the HPS header to the main and start-up air extraction vacuum pumps. In addition, there may be other HPS consumers as per plant or external steam consumer requirements.

Low-Pressure Steam System

General The purpose of the low-pressure steam (LPS) system is to deliver LPS to the in- and out-of-plant consumers as required.

Hitachi Zosen Inova Technology. Figure 42
LP steam header

Description The LPS system is typically set at approx. 3–5 bars (45–75 psi) pressure. The system supplies steam to the feed water tank with deaerator, to the air preheater and possibly to external LPS consumers, such as a district heating system.

If the turbine is in operation, LP steam is taken from the turbine extraction point to the LPS header. If the turbine is not in operation, the LPS is supplied from the HPS header via the HPS/LPS reduction station. The LPS system consists of an LPS header (Fig. 42), piping from the turbine extraction to the LPS header, piping from the HPS header via the steam pressure reduction station to the LPS header, and piping from the LPS header to the internal users.

Condensate System

General The purpose of the condensate system is to collect and treat all condensates which arise in the plant. The more clean condensates can be recovered, the less demineralized water has to be made up.

Description The condensate system consists of a main condensate tank, the redundant condensate pumps, as well as the connection to all condensate sources. These are typically the main condenser via its condensate pumps, steam ejector condenser, as well as all condensate traps of the internal steam systems and heaters. To recover additional energy contained in the condensate from HPS systems and heaters, high-pressure condensate is first flashed and the generated LPS is utilized in the plant.

Demineralized Water System

General The purpose of the demineralized water system is to clean the available raw water to obtain the required purity and chemistry required for a steam boiler and turbine system in order to prevent corrosion and equipment damage. Additional water purification steps may be needed if the raw water source for the EfW plant is of low quality.

Description A demineralized water system as shown in Fig. 43 is designed for a continuous supply of demineralized water for the boiler. It is designed with sufficient capacity and redundancy. This includes redundancy of all components with the exception of piping and storage tanks.

The typical demineralized water system consists of a raw water tank, a cation exchanger, an anion exchanger, a regeneration unit, a neutralization unit, a demineralized water storage tank, and demineralized water pumps.

Nitrogen Oxides (NO_x) Reduction Measures

General Nitrogen oxides (NO_x, mainly consisting of NO and NO_2) reduction includes multiple measures with the goal to have lowest concentrations of nitrogen oxides in the flue gases leaving the plant. Nitrogen in the combustion air and even more nitrogen chemically bound in the waste will partially react with the oxygen at combustion temperatures and form NO_x.

Hitachi Zosen Inova Technology. Figure 43
Demineralized waste system equipment

Primary measures have the goal to minimize the formation of NO_x. The flue gas recirculation presented above is a powerful means to achieve this goal. Generally the reduction of the oxygen level in the flue gases and in particular the oxygen level achievable using flue gas recirculation has resulted in uncontrolled NO_x level in Hitachi Zosen Inova's latest plants of only about half the levels of plants built in the 1970s.

Secondary measures have the goal to reduce NO_x which has already been formed during combustion. The most known measures include a reaction of ammonia with the NO_x to form nitrogen and water vapor.

$$2NO + 2NH_3 + O_2 \rightarrow 2N_2 + 3H_2O$$

$$2NO_2 + 4NH_3 + O_2 \rightarrow 3N_2 + 6H_2O$$

This reaction can technically occur at different temperature ranges:

- At high temperatures, as a spontaneous and selective reaction, referred to as the selective non-catalytic reduction (SNCR)
- At lower temperatures, only with the aid of a catalyst, referred to as the selective catalytic reduction (SCR)

Selective Non-catalytic Reduction (SNCR)

Overview The SNCR process is based on the principle of injecting an ammonia-containing reagent into the furnace where the temperature range is between 850°C (1,560°F) and 950°C (1,740°F). The reagent is typically an aqueous ammonia or urea solution. The required temperature range exists in the upper part of the furnace (the first pass of the boiler). However, the actual furnace temperature at any given injection point may vary a lot due to typically varying combustion conditions on the grate caused by highly varying waste properties, by load changes, or by fouling of the boiler walls. If actual temperatures become too low, the reaction is not efficient any more and the ammonia slip at the stack is increased. If actual temperatures become too high, the following secondary reactions become significant and actually ammonia is decomposed or burned leading to highly increased reagent consumption and NO_x emissions.

$$4NH_3 + 5O_2 \rightarrow 4NO + 6H_2O \text{ at high temperatures}$$

$$4NH_3 + 3O_2 \rightarrow 2N_2 + 6H_2O$$

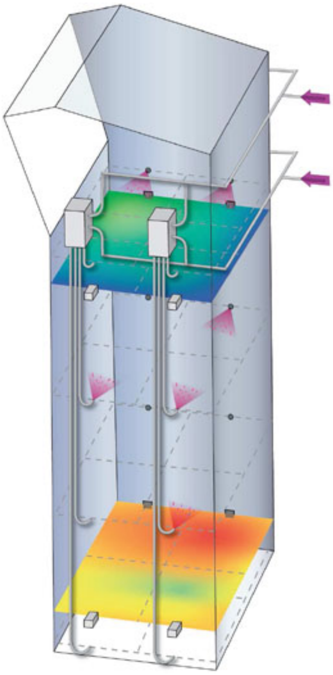
The key of a good SNCR process is to inject the reagent at exactly the right temperature level and to mix it evenly into the flue gases.

The Dynamic NO_x Reduction (DyNOR®) System

CFD calculations show that flue gases do not mix well perpendicularly to the flow direction [4]. Therefore, the conditions in a vertical segment of the first boiler pass are predominantly defined by what happens on the grate at the bottom of that segment. And since the waste is so inhomogeneous, this can mean that at the same location the short-term temperature changes can be 100°C (180°F) and more.

The main features of the Hitachi Zosen Inova DyNOR® system (Fig. 44) are the following:

- The first boiler pass is divided into several vertical segments, depending on the size of the boiler.
- Three or four injection levels are determined and on each level one or more injection nozzles are dedicated to each segment.
- In each vertical segment, a fast infrared temperature sensor measures the instantaneous temperature.
- For each vertical segment, the controller calculates the optimum injection level at each point in time.
- For each vertical segment, a fast distributor directs the reagent and the atomizing medium without time lag to the nozzles closest to the optimum injection level.



Hitachi Zosen Inova Technology. Figure 44
DyNOR® setup

As with conventional SNCR systems, the configuration of the nozzles is important to cover the entire cross section of the boiler or segment and assure an even distribution of the reagent across the whole cross section. Other measures that improve the performance of the SNCR system are:

- Keeping steady combustion conditions on the grate with a high-quality combustion control system automatically adjusting to the varying qualities of the waste incinerated (see section "Combustion Control System")
- Good distribution and mixing of secondary air and recirculated flue gases which level out oxygen and temperature profiles as achieved by a swirl flow pattern (see section "Flow Optimization with Swirl Injection")
- No unnecessary internals in the SNCR reaction zone which would disturb the flow pattern and therefore compromise the even distribution of reagent

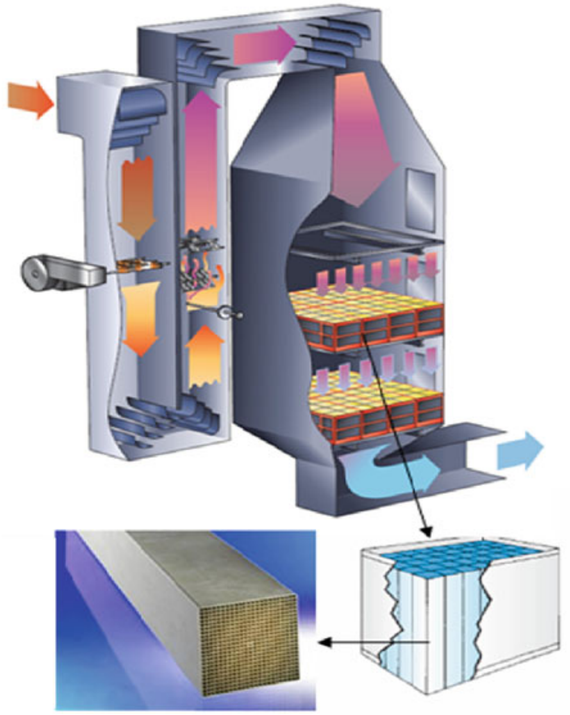
Selective Catalytic Reduction (SCR)

Overview In the SCR method, the conversion of nitrogen oxides is assisted by a catalyst and can take place at temperatures between about 180–350°C (350–650°F). The reaction is up to typically 90% efficient without stoichiometric overconsumption of ammonia, and thus only with minimum ammonia slip at the stack. As a secondary benefit, dioxins and furans are also oxidized on the catalyst. An undesired secondary reaction is the conversion of SO_2 to SO_3, which can then further react with ammonia to form solid ammonium sulfate salts which deactivate the catalyst. The reaction will also happen at lower or higher temperatures, but secondary effects such as catalyst deactivation and scaling or energy efficiency considerations have practically led to the mentioned temperature range.

At operating temperatures below about 220°C (430°F), the catalyst needs to be periodically regenerated, during operation. In order to evaporate possible salt deposits from the surface of the catalyst modules, the flue gas entering the catalyst is heated up to about 320°C (610°F) over a period of about 6 h.

To implement an SCR process (Fig. 45) in an EfW plant, the following components are typically needed:

Gas–Gas Heat Exchanger A gas–gas heat exchanger is only needed in tail-end configuration. The flue gases entering the catalyst have to be reheated. The most economic way to reheat flue gases is by cooling down flue gases coming from the catalyst by means of a gas–gas heat exchanger. This can be done by a plate or tube-type steel heat exchanger. An external heat source, preferably a steam preheater, then only has to compensate the gas–gas heat exchanger temperature driving force.



Hitachi Zosen Inova Technology. Figure 45
SCR reactor with layers, catalyst module, and element

If flue gases have to be reheated after a wet scrubber, the cold heat exchanger surfaces have to be protected against corrosion.

Steam Preheater Final preheating can be done with a steam–gas heat exchanger. Steam from the boiler steam drum is typically between 40 and 60 bar (600–900 psi) which allows to heat flue gases to a temperature of about 220–250°C (430–480°F). The steam/gas preheater is a self-supported, flue gas-tight, welded structure made up of sheet-metal sections and equipped with peripheral frame into which the steam bundles can be inserted.

Duct Burner If steam is not available or if the flue gases have to be heated above the temperature range that is practical for a steam preheater, for example, for regeneration, the gas is heated with a duct burner.

Ammonia Injection and Distribution Ammonia-containing gas or solution is injected into flue gas duct upstream of the catalyst. The turbulence generated in a static mixer installed downstream the ammonia injection assures a good distribution of the ammonia in the flue gas and a homogeneous velocity, temperature, and concentration profile. Since in the catalyst there is virtually no further mixing, best reaction efficiency and lowest ammonia slip is heavily dependent on the conditions prior to entering the catalyst.

Catalyst Reactor The reactor is made as a self-supported, flue gas-tight, welded, rectangular steel casing with provisions to hold the catalyst modules. Within the reactor, the catalyst is typically arranged in 1–3 layers and space is provided for a spare layer.

The catalyst modules are ready-to-install steel cases containing the catalyst elements. Depending on the cross section of the reactor, a layer contains several modules. The height of one layer is typically 0.5–1 m (1.5–3 ft). Between the layers there is about 2 m space for access. Depending on the possible dust load in the flue gases, sootblowers are installed upstream of each layer (Fig. 45).

Typically, a plate or honeycomb-type catalyst consisting of TiO_2 (titanium dioxide) support material plus V_2O_5 (vanadium pentoxide) and WO_3 (tungsten trioxide) active substances is used. The channel pitch of a typical SCR catalyst is only a few millimeters and is one of the important design criteria. The smaller the pitch the more surface there is per catalyst volume, however, the bigger is the risk of plugging.

Acid Gas Removal

General Acid gases such as hydrogen chloride, hydrogen fluoride, sulfur dioxide, and sulfur trioxide are formed when wastes containing chlorine, fluorine, and sulfur are burned. Removal of acid gases involves the use of an alkaline reagent, which reacts with the acid and forms a nonvolatile salt, which in most cases has to be disposed of. With the example of hydrated lime as the reagent, the following simplified reactions take place:

$$Ca(OH)_2 + H_2O + SO_2 = CaSO_3 + 2H_2O$$

$$Ca(OH)_2 + H_2O + SO3 = CaSO_3 + 2H_2O$$

$$CaSO_3 + 1/2O_2 = CaSO_4$$

$$Ca(OH)_2 + 2HCl = CaCl_2 + 2H_2O$$

$$Ca(OH)_2 + 2HF = CaF_2 + 2H_2O$$

$$Ca(OH)_2 + 2HF = CaF_2 + 2H_2O$$

$$Ca(OH)_2 + CO_2 = CaCO_3 + H_2O \text{ (unwanted)}$$

Depending on the physical state in which the reaction takes place, three main types of processes are differentiated:

- Wet: The flue gases are cooled down to saturation temperature in the process and the reaction takes place in an aqueous solution or slurry of the reagent.
- Semi-wet or semidry: Semi-wet means the reagent is injected as an aqueous solution or slurry. Semidry means the reagent is injected as a dry powder, but is subsequently conditioned by a separate water injection. However, in both cases, the water evaporates in the process so the reaction products are removed as a dry powder while the water vapor stays in the flue gases. The flue gas temperature is controlled in the process by the amount of water added.
- Dry: The reagent is injected and the reaction products are removed as a dry powder. Temperature control happens upstream of the process.

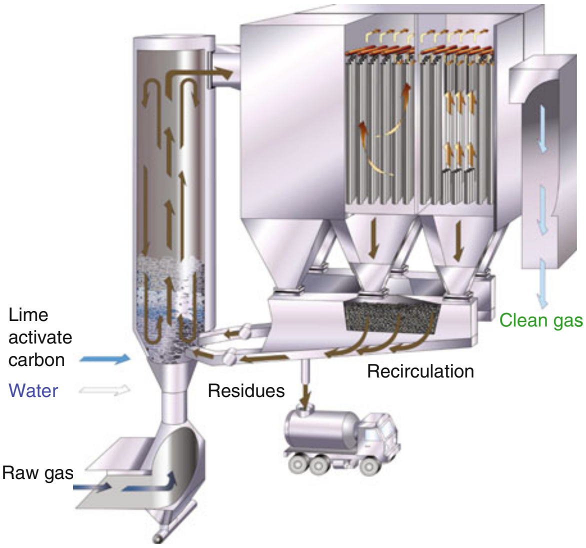
While Hitachi Zosen Inova offers all the above processes and even combinations of such processes, the following detailed description focuses on the HZ Inova Semidry FGT process, which is one of the most efficient and widely used acid gas removal processes in the EfW sector.

Description of the HZ Inova Semidry FGT Process

Process The flue gases are brought into intensive contact with the reagents in a fluidized bed reactor. Unused reagents and reaction products entained with the flue gases from the reactor and collected in the downstream fabric filter. A large portion of the solids removed in the filter are recirculated and reinjected into the reactor in order to optimize the utilization of the reagents. A small amount is discharged.

The HZ Inova Semidry FGT process (Fig. 46) has the following advantages:

- Highly efficient reaction due to high turbulence and good mixing in the reactor and conditioning water addition to the reaction zone
- High flexibility relative to load changes and contaminant peaks due to the large amount of reagent in contact with the flue gas at any time
- Optimized use of reagent due to solids recirculation
- High availability due to simple construction and no moving parts



Hitachi Zosen Inova Technology. Figure 46
Hitachi Zosen Semidry FGT process schematic

Process Control The HZ Inova Semidry FGT process has three control loops, which regulate the process to achieve low emissions and minimize the consumption of hydrated lime.

The first loop continuously controls the pressure drop across the fluidized bed and maintains fluidized bed level constant by adjusting the flow of recirculated solids.

The second control loop controls the flue gas temperature at the outlet of the reactor by regulating water injection at the bottom of the fluidized bed (Fig. 47).

The third loop controls the acid gas emissions at the stack by regulating the flow of fresh hydrated lime to the reactor (Fig. 48).

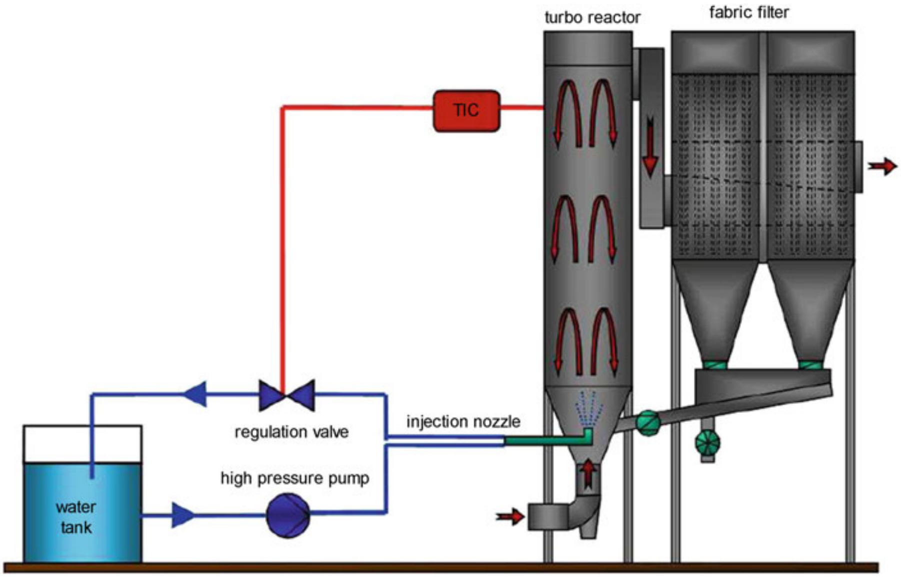
The HZ Inova Semidry FGT process is automatically controlled by the plant DCS including sequences for start-up, normal shutdown, and emergency shutdown.

HZ Inova Semidry Reactor The flue gas enters the reactor centrally from below through a venturi-type restriction. Therein, the flue gas velocity is high enough to ensure that all solids are entrained into the reactor or prevented from falling down against the flue gas flow. The fluidized bed is sustained in the reactor by larger particles that are kept in suspension by the upward flowing flue gas. Smaller particles and the flue gases exit the reactor laterally at top of the reactor. Depending on the flow pattern formed at the top of the reactor, a part of the solids will sink back to the bed along the wall where the velocities are lower. The flue gas has a mean residence time of approx. 3 s in the reactor.

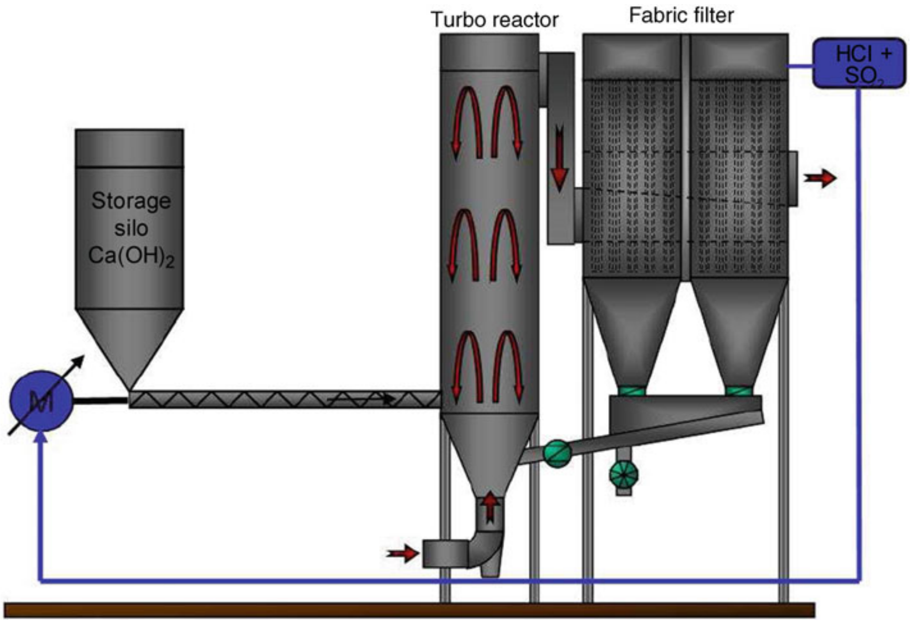
The injection ports for hydrated lime, for other reagents, for the recirculated solids, and for the water are located in the conical inlet of the reactor.

Recirculation Conveyor In order to form a fluidized bed and to fully utilize the reagents, the largest part of the solids separated inside the fabric filter is circulated back to the reactor via a moving bed conveyor as shown in Fig. 49. Fluidizing air from special blowers is injected through the bottom of the conveyors. This ensures the steady flux and keeps the solids in a permanent motion thus preventing the formation of deposits.

Water Injection Lance The water is injected through a lance with a specially developed hydraulic nozzle which atomizes the water to the required droplet size of about 30–100 µm. The nozzles are equipped with an automatic cleaning system consisting of a rubber cap which can be bloated periodically with pressurized air in order to break off possible deposits (Fig. 50).



Hitachi Zosen Inova Technology. Figure 47
Temperature control loop



Hitachi Zosen Inova Technology. Figure 48
Emission control loop

Hitachi Zosen Inova Technology. Figure 49
Reagent recirculation conveyor and residue discharge

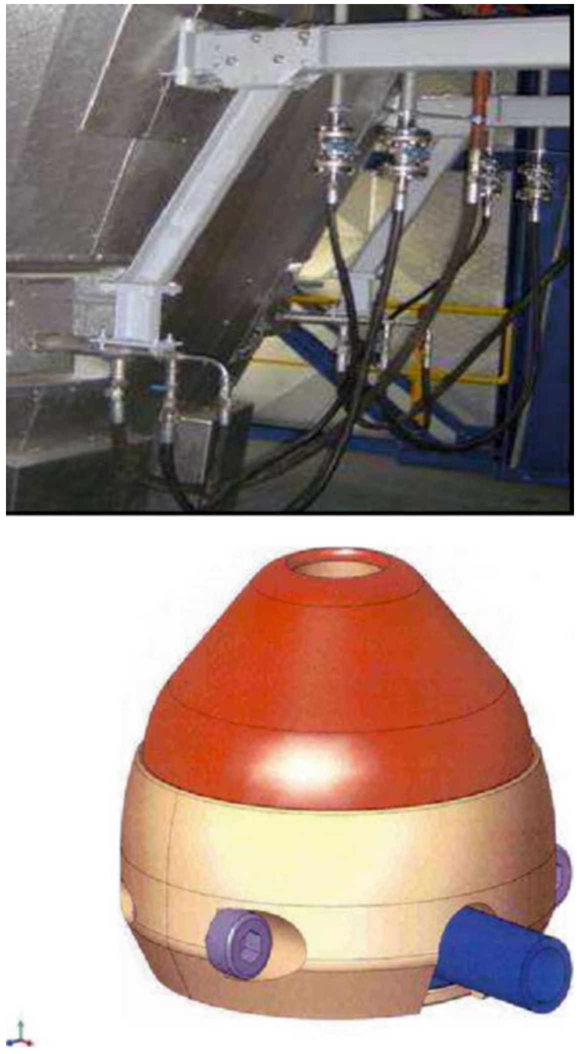
High-Pressure Water System Water for injection is supplied from the head tank via a high-pressure pump at about 30–40 bar (450–600 psi). An example of the HP water system equipment is shown in Fig. 51. The amount of injected water is controlled by a valve in the return line to the head tank.

As long as the water does not contain solids incompatible with the injection nozzle, any type of water can be injected. This is often used to safely dispose of contact or wash water collected during such operations. This allows EfW plants to be effluent free.

Volatile Heavy Metal and Organic Component Removal in the HZ Inova Semidry FGT Process Apart from the chemical reactions between the reagent and the acid gases, the immense surface of the large amount of reagent particles also serves to adsorb other contaminants. In order to enhance the adsorption of volatile heavy metals, predominantly mercury, and volatile organics, such as dioxins and furans, activated carbon or lignite coke is also injected into the reactor. Only small amounts of such an additive assure, due to the good mixing and the high solids recirculation rate in the HZ Inova Semidry reactor, that most such contaminants are safely removed.

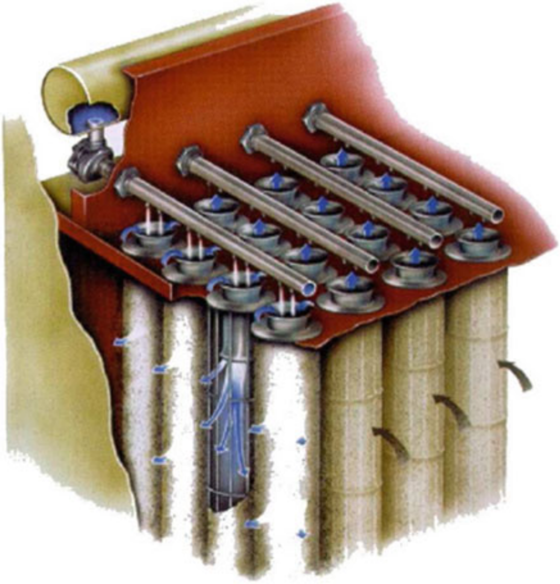
Fabric Filter

General The purpose of the fabric filter (Fig. 52) is to separate the ash particles in the flue gas as well as the reaction products from the HZ Inova Semidry reactor from the flue gas.



Hitachi Zosen Inova Technology. Figure 50
Water injection lance and nozzle

Hitachi Zosen Inova Technology. Figure 51
HP water skid



Hitachi Zosen Inova Technology. Figure 52
Filter chamber

Description The fabric filters are used for the separation of solids from gases. The separation is a physical process where the solids are filtered on the surface of a gas-permeable fabric. Due to the intensive contact of the flue gas and the adsorbents in the filter layer, the removal of pollutants from the flue gas is also further improved.

The fabric filter consists of several filter chambers into which the total flue gas stream is divided up. Each chamber can be isolated from the flue gas stream in case of a leakage in the filter material. The flue gases enter into the respective filter chamber and are deflected by a distributor panel to assure a uniform distribution of the gases through the filter and separate a large part of the solids carried with the flue gas. The gases flow around the outside of an array of the filter bags suspended in the chamber. Then they pass through the filter media, which are supported on a wire cage, and are collected from the inside of the filter bags through the top of the filter chamber into a common clean gas duct.

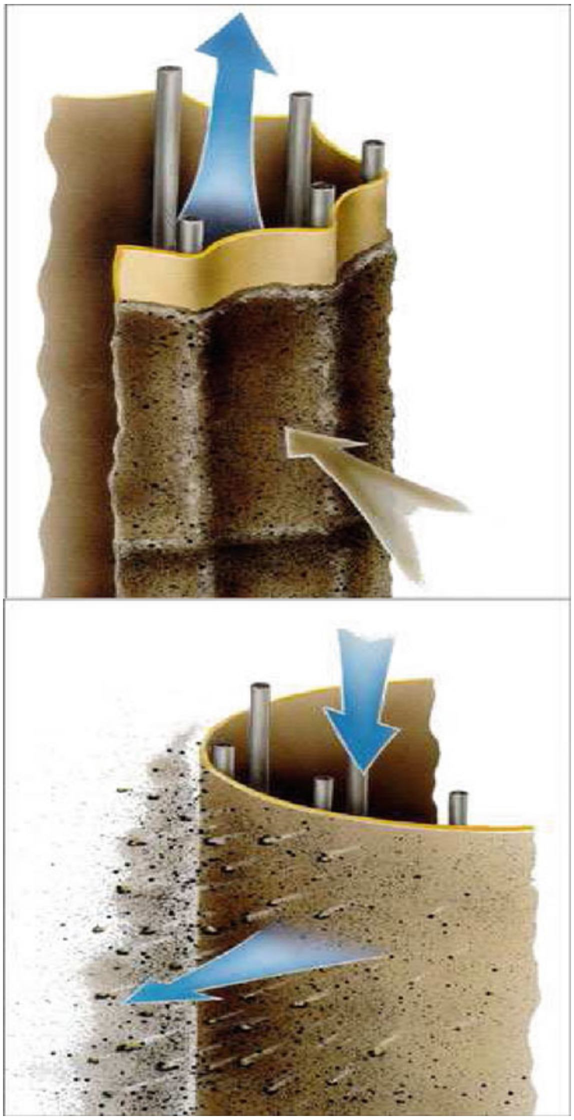
Due to the flow characteristics induced by the chamber geometry as well as the pressure drop, the flue gas passes uniformly through the filter bags. This causes a uniform deposition of the solids on the outside of each individual filter bag. These need to be cleaned when solids layer on the bags becomes too thick and the pressure drop across the filter becomes too large (Fig. 53). This happens in an automated process. The solids are purged by suddenly inflating the bags with pulses of compressed air from the clean gas side. The filter bags are inflated for a short time which breaks off the solid layer on the outside. These fall down into the hoppers below the filter bags, from where they are transported away by the recirculation conveyor. The hoppers of the fabric filter can be isolated from the pneumatic conveyors by means of manually operated slide gate valves.

The cleaning air is supplied via compressed air reservoirs (one per filter chamber). From there it is distributed by a series of nozzle lances to each row of filter bags.

The fabric filter is equipped with differential pressure, temperature, and level-measuring instruments. The devices assure the supervision of the filter operation and the detection of possible malfunctions.

On top of the clean gas chamber, there is an accessible platform with complete enclosure – the filter penthouse. It is dimensioned such that there is enough room to install and remove bags from the fabric filter via walk-in plenum to the clean gas chamber.

In order to prevent deposits of and corrosion by the hygroscopic and corrosive reaction products at cold temperatures, the fabric filter hoppers are heated with electric heaters during start-up and shutdown.



Hitachi Zosen Inova Technology. Figure 53
Filter bag during operation and cleaning

Additive Systems

General The purpose of the additive feed system as shown in Fig. 54 is to store and transport additive (reagents and sorbents) to the HZ Inova Semidry reactor.

Reagent System (Hydrated Lime) The reagent silo typically contains a few days' to a weeks' supply. The discharge from the silo is ensured by a mechanical discharge and dosing system. Then the hydrated lime is conveyed to the reactor pneumatically and injected as dry powder together with the transport air directly into the HZ Inova Semidry reactor. Solids dosing and transport systems are preferably designed redundant since they assure compliance with the plant's emission limits.

The concentrations before and after the flue gas treatment system of the lead acid gas component are used to

Hitachi Zosen Inova Technology. Figure 54
Reagent and residue silos

control the flow of hydrated lime. The hydrated lime dosing and transport system is dimensioned such that also expected peaks of pollutants can be safely treated.

Sorbent System (Activated Carbon) The sorbent storage can be a silo or a big bag, depending on the size of the plant. It typically contains a few days' to a weeks' supply. The discharge from the storage is ensured by a mechanical discharge and dosing system. Then the activated carbon is conveyed to the reactor pneumatically and injected as dry powder together with the transport air directly into the HZ Inova Semi-dry reactor. Solids dosing and transport systems are preferably designed redundant since they assure compliance with the plant's emission limits.

Hitachi Zosen Inova Technology. Figure 55
ID-fan with thermal and noise insulation

The amount of activated carbon to be injected into the reactor depends on the expected concentrations of volatile organics (dioxins/furans) as well as volatile heavy metals (mercury). The quantity of activated carbon per volume of flue gas is set during commissioning and annual emission testing.

Induced Draft Fan

General The purpose of the induced draft fan (ID-fan, Fig. 55) is to maintain a negative draft in the combustion chamber and flue gas cleaning system and then to push the flue gas through the stack.

Description The ID-fan is controlled to keep the furnace pressure constantly a little below atmospheric. Since the ID-fan is the largest electrical consumer in the plant, this is done by means of a variable speed drive which gives the best energy efficiency for this task.

The fan is equipped with openings for inspection and cleaning purposes. The fan and its drive are mounted on a common concrete frame structure designed to absorb and dampen vibrations. Bearing temperature and vibrations are constantly monitored.

Emissions: Continuous Emission Monitoring System

General The purpose of the continuous emission monitoring system (CEMS) is to monitor purity of the flue gas exiting the stack and to provide signals to the various control loops of the air pollution control systems.

Flue Gas Quality The following table shows European and US standard emission limits as well as a range of typical values achieved in various Hitachi Zosen Inova plants (Table 1).

While all legal emission limits can be safely kept, it is also possible to design plants which lead to substantially lower emissions as required by the referenced national limits. However, this is often a factor of the waste composition and of the complexity of the air pollution control equipment – thus of the cost which a region is ready to spend for responsible waste treatment.

Hitachi Zosen Inova Technology. Table 1 Typical Emission Limits and Hitachi Zosen Inova EfW Plant emission data

	EU		USA		Hitachi Zosen INOVA	
	(1) (3)	(2)	(1)	(2) (4)	(1) (13)	(2) (13)
CO	50 mg/Nm^3	56 ppmv	89 mg/Nm^3	100 ppmv	10–50	11–56
TOC (5)	10 mg/Nm^3	13 mg/scm	–	–	1–10	1.3–13
Particulate	10 mg/Nm^3	13 mg/scm	15 mg/Nm^3	20 mg/scm	1–10	1.3–13
HCl (6)	10 mg/Nm^3	8.6 ppmv	29 mg/Nm^3	25 ppmv	1–10	1.3–13
HF	1 mg/Nm^3	0.9 ppmv	–	–	0.1–1	0.09–0.9
SO_2 (7)	50 mg/Nm^3	24.5 ppmv	61 mg/Nm^3	30 ppmv	5–50	2.5–24.5
NO_x as NO_2 (8)	200 mg/Nm^3	136 ppmv	220 mg/Nm^3	150 ppmv	40–200	27–136
NH_3 (8)	10 mg/Nm^3	18.5 ppmv	–	–	1–10	1.8–18
Hg (9)	0.05 mg/Nm^3	0.065 mg/scm	0.038 mg/Nm^3	0.05 mg/scm	$<$0.038	$<$0.05
Cd (+Tl) (10)	0.05 mg/Nm^3	0.065 mg/scm	0.008 mg/Nm^3	0.01 mg/scm	$<$0.008	$<$0.01
Pb (+HM) (11)	0.5 mg/Nm^3	0.65 mg/scm	0.107 mg/Nm^3	0.14 mg/scm	$<$0.1	$<$0.14
PCDD + PCDF	0.1 ng/Nm^3 TEQ		0.1 ng/Nm^3 TEQ	13 ng/scm 12	0.01–0.1	1–10

(1) European emission values are daily average values or average values over a sampling period based on 11% O_2 in dry flue gas standardized at 273K and 101.3 kPa
(2) US emission values are based on 7% O_2 in dry flue gas standardized at
(3) As per EU Directive 2000/76/EC [5]
(4) As per EPA Title 40 Part 60, Subpart Eb [6]
(5) Total organic carbon (TOC)
(6) While the EU value is absolute, the US value can be exceeded if the removal efficiency is at least 95%
(7) While the EU value is absolute, the US value can be exceeded if the removal efficiency is at least 80%
(8) Local air permits in the USA often reduce NO_x and introduce NH_3 emission limits, specially in non-attainment zones
(9) While the EU value is absolute, the US value can be exceeded if the removal efficiency is at least 85%
(10) The EU value includes the sum of Cd + Tl, the US value only Cd, which is however by far the largest portion of the sum
(11) The EU value includes the sum of the following heavy metals Sb + As + Pb + Cr + Co + Cu + Mn + Ni + V, the US value is only for Pb, which is one of the larger portions of the sum
(12) The EU value is based on a toxic equivalence concept which calculates a weighted sum of all different PCDD and PDCF isomers is calculated. The US value uses the sum of all isomers without weighing. For typical mixtures of PCDD and PCDF isomers present in flue gases of EfW plants, these emission limits are similar
(13) Same units as used in preceding columns

Description of the CEMS A separate CEMS monitors the flue gas properties and composition at the stack for each incineration line. Each system mainly consists of:

- Temperature, pressure, and flow rate measurement
- Particulate measurement
- Complete extractive gas analyzer for concentration measurement of typically the following flue gas components: H_2O, O_2, CO, HCl, SO_2, NO_x, NH_3, VOC.
- Computer-based data evaluation and storage system

The instruments are directly installed on either the discharge ductwork between the ID-fan and stack or the stack. The flue gas is extracted through a heated extraction line and conveyed to the analytical instruments installed in the CEMS container.

The emission measurement system is designed to meet the particular requirements of the operating permit for the plant. It complies with the applicable directives for installation and quality assurance and has the possibility to generate direct reports for the supervising Environmental Protection Agency.

Stack

General The purpose of the stack is to expel the clean flue gases to the atmosphere at a high enough point to ensure adequate dispersion (Fig. 56).

Hitachi Zosen Inova Technology. Figure 56
Stack with individual support of single flues (Moerdijk, the Netherlands)

Description The cleaned flue gases enter the stack with a temperature of about 150°C (300°F). In multiple train plants, all flues can be integrated into the same stack structure. The internal flues are made of steel for flue gases above typically 130°C (270°F) or of fiber-reinforced plastic below. The flues are thermally insulated to avoid excessive condensation along the walls. The stack support structure can be steel or concrete.

Underneath the flue gas inlet to the stack, condensed water vapor or rain water is collected on an inclined stack base plate and discharged via a condensate discharge line. The condensate discharge is drained into the contact water storage.

The stack includes a safety ladder to a maintenance and work platform at the top of the stack from where the flues can be monitored and the lightning protection and air traffic lighting systems can be maintained. Depending on the location of emission monitoring nozzles an additional platform may be located in the middle of the stack.

Since the stack is normally the most visible part of an EfW plant, there are ideas how a stack could be converted to a landmark, for example in Fig. 57.

Future Directions

The EfW (or WTE) technology described above has been shown to be reliable, efficient, and proven:

- Reliable means that a large number of plant owners operate their Hitachi Zosen Inova plants in the range of 7,900–8,300 h per year, which corresponds to 90–95% availability.
- Efficient means that all national emission limits can be met, sometimes with a considerable margin. It also means that energy can be efficiently recovered in the form of electricity and/or heat meeting or exceeding the relevant energy efficiency standards [7].
- Proven means that this technology has been developed and has matured with the over 500 reference plants built worldwide since the 1930s.

It is evident that the Hitachi Zosen Inova technology is mature. Technical developments concentrate mainly on the following areas:

- Increase energy efficiency and electricity production og EfW plants

Hitachi Zosen Inova Technology. Figure 57
Stack with combined support of flues and architectural treatment (Osaka, Japan)

- Improve the automated control of the EfW plants in order to further increase average plant capacity and availability.
- Improve monitoring of plant components in order to allow operators to track the plant efficiency and better plan preventive maintenance actions.
- Further improve stack emission limits.
- Further improve bottom ash quality and recycling options for bottom ash (recovery of valuable resources such as copper, aluminium, brass and other metals).

However, since EfW technology in many areas still commercially competes with relatively cheap landfilling or is not available for large economic areas, one goal of development is always concentrated on plant economics:

- Achieve the same performance with less and simpler equipment
- Find less-expensive materials of construction, in terms of investment and maintenance cost
- Further improve the operation flexibility and availability of the plant thus reducing O&M cost

Last but not least, the goal must be to do all that is necessary to make EfW plants politically acceptable and a part of the infrastructure in the densely populated areas. Wastes are created and energy is needed where the people live. EfW plants are an essential part needed to close the loop of the material cycle [3]. Whatever cannot be reused or recycled can be efficiently treated for energy recovery in an EfW plant. An excellent example is the Isséane EfW plant that started operations in 2008 in France:

- It is located in the middle of the Paris metropolitan area about 5 km (3 miles) from the Eiffel Tower along the river Seine.
- It is an integrated part of the city’s recycling activities and contains a plastics sorting and recycling facility.
- It serves a population of about one million people all living within a short distance from the plant.
- It provides electricity to the national electrical grid.
- It provides heat to the city’s district heating system.
- However, both electricity and heat are consumed in the vicinity of the plant. This concept helps reduce the need for long-distant transport of energy and the associated complexity of energy distribution.
- The EfW plant was designed to fit architecturally into the neighborhood and, in fact, may not even be recognized as an EfW plant.

Although the total cost of the Isséane EfW plant (Fig. 58) is on the high side compared to other modern EfW plants, it is an excellent example of sustainability within the material and energy cycles. Research and development is focusing on all aspects needed to make such a concept available to more and more highly populated areas. This includes emissions and architecture to increase acceptability to the public as well as efficiency, reliability, and total cost in order to increase its affordability to communities in developing nations.

Hitachi Zosen Inova Technology. Figure 58
Isséane energy-from-waste plant, Paris, France

Bibliography

1. Kjeldsen P, Barlaz MA, Rooker AP, Baun A, Ledin A, Christensen TH (2002) Present and long-term composition of MSW landfill leachate: a review. Critical Reviews in Environmental Science and Technology 32(4):297–336
2. Kaufman SM, Themelis NJ (2009) Using a direct method to characterize and measure flows of municipal solid waste in the United States. J Air Waste Manage Assoc 59:1386–1390. doi:10.3155/1047-3289.59.12.1386, ISSN:1047-3289; Copyright 2009 Air & Waste Management Association, Earth Engineering Center, Columbia University, New York
3. European Environment Agency (2007) Sustainable consumption and production. In: State of the environment report no 1/2007 (Chap 6). Office for Official Publications of the European Communities, Luxembourg. http://www.eea.europa.eu/publications/state_of_environment_report_2007_1
4. Sigg F, Halter R, Chromec P (2010) DyNOR® DeNOx performance confirmed in further MSW plants, NAWTEC18-3504. In: Proceedings of the 18th annual North American waste-to-energy conference, Orlando. Copyright 2010 by ASME
5. Directive 2000/76/EC of the European Parliament and of the Council of 4 December 2000 on the Incineration of Waste, 28.12.2000, Official Journal of the European Communities, L 332/91
6. US Electronic Code of Federal Regulations (Title 40, Part 60, Subpart Eb) current as of 29 July 2010
7. Reimann DO (2004–2007) Results of specific data for energy, R1 plant efficiency factor and net calorific value (NCV) of 231 European WtE plants. In: CEWEP Energy Report II (Status 2004–2007). http://www.cewep.eu/studies/climate-protection/index.html

HIV/AIDS Global Epidemic

Phyllis J. Kanki
Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA, USA

Article Outline

Glossary
Definition of the Problem
Introduction
Basic Virology of HIV

HIV and the Cause of AIDS
Epidemiology of the Global HIV Pandemic
Sub-Saharan Africa
Asia
Eastern Europe and Central Asia
Caribbean
Central and South America
North America and Western and Central Europe
Middle East and North Africa
Oceania
Prevention
Genetic Diversity of HIV
Treatment
Vaccine
Future Directions
Bibliography

Glossary

Acquired immunodeficiency syndrome (AIDS) A clinical syndrome caused by the human immunodeficiency virus (HIV). Its pathogenesis is related to a qualitative and quantitative impairment of the immune system, particularly a reduction of the CD4+ helper T lymphocyte cell count (surrogate marker of the disease). After an average of 10 years, if untreated, HIV + individuals can develop opportunistic diseases (i.e., infections and cancers rarely detected in people with normal immune systems). The natural history of the disease can be dramatically modified with administration of combination therapy composed of antiretroviral (ARV) drugs.

CCR-5 A cell membrane protein expressed on several cell types including peripheral blood-derived dendritic cells, CD34+ hematopoietic progenitor cells, and certain activated/memory Th1 lymphocytes. This receptor is well defined as a major coreceptor in conjunction with CD4+, implicated in susceptibility to HIV-1 infection.

CD4+ A large glycoprotein that is found on the surface of helper T lymphocyte cells, regulatory T cells, monocytes, and dendritic cells. Its natural function is as a coreceptor that assists the T cell receptor (TCR) to activate its T cell following an interaction with an antigen-presenting cell. CD4+ is a primary receptor used by HIV-1 to gain entry into host T cells.

CD4+ T cell An immune cell, lymphocyte (white blood cell) characterized by the CD4+ antigen (protein) on its surface. This is a T lymphocyte considered to have a “helper” function to enhance the cellular immune response. The CD4+ is the primary receptor for the HIV virus, and upon infection, the virus can destroy the CD4+ cell. In HIV-infected people, the drop in CD4+ T lymphocyte cells is a major determinant of the progression of HIV infection to AIDS.

Coreceptor (CCR-5 or CXCR-4) Protein molecules on the surface of lymphocytes or monocytes that bind to the gp120 protein of HIV and facilitate, with CD4, binding, fusion, and entry of the virus into the susceptible cell.

CXCR-4 An alpha-chemokine receptor specific for stromal-derived factor-1 (SDF-1 also called CXCL12), a molecule endowed with potent chemotactic activity for lymphocytes. This coreceptor is one of several chemokine receptors that HIV isolates can use to specifically infect CD4+ T cells.

DNA (deoxyribonucleic acid) A nucleic acid that contains the molecular basis of heredity for all known living organisms and some viruses and is found in the nuclei and mitochondria of eukaryotes. Chemically, DNA consists of two polymer strands of units called nucleotides made up of one of four possible bases plus sugar and phosphate groups. The polymers are joined at the bases by hydrogen bonds to form a double helix structure.

Fusion of virus and cell membranes A merging of cell and virus membranes that permits HIV proteins and nucleic acids to enter the host cell.

Fusion/entry inhibitors A class of ART drugs that interferes with the virus’ ability to fuse with the target cell’s outer membrane, thereby blocking entry of the HIV into the host cell.

gp120 The major HIV envelope glycoprotein having a molecular weight of 120 that protrudes from the outer surface of the virion. This glycoprotein binds to a CD4+ receptor on a T cell to facilitate entry of the virus into the cell.

Human immunodeficiency virus (HIV) The virus that causes acquired immunodeficiency syndrome (AIDS). It is a lentivirus belonging to *Retroviridae* family and was discovered in 1983 by Robert Gallo and Luc Montagnier. HIV infects and destroys helper T cells of the immune system causing a marked reduction in their numbers. Loss of CD4 cells leads to generalized failure of the immune system and susceptibility to life-threatening opportunistic infections. It is transmitted mainly through sexual intercourse, exchange of contaminated syringes among intravenous drug users, and contaminated blood transfusion. HIV-1 is the HIV type most frequently detected HIV worldwide and responsible for the global pandemic.

HIV-1 subtypes or clades Genetically related HIV strains that are essentially phylogenetically equidistant, generating a starlike phylogeny. Subtypes A, B, C, D, F, G, H, J, and K are currently known subtypes A, B, C, and D are highly prevalent others have low prevalence and limited geographic distributions.

HIV-2 The second HIV virus discovered in West Africa in 1984, the virus is more closely related to the simian immunodeficiency virus of primates. Although HIV-2 can cause AIDS, it has a distinct epidemiology, lower rate of transmission, and slower progression to disease.

Incidence Rate describing the number of new cases of disease occurring within a given time period, expressed as new cases per person-time.

Integrase An enzyme found in retroviruses including HIV that permits the reverse transcribed viral DNA to be integrated into the infected cell's DNA. Integrase is an enzyme encoded by the polymerase gene of HIV.

Integrase inhibitors A class of ART drugs that blocks the viral integrase the enzyme HIV uses to integrate its genetic material into its target host cell DNA.

Nucleus A membrane-enclosed central compartment of a cell that functions to contain the genomic DNA and to regulate gene expression.

Prevalence Number of cases of disease in a defined population at a specific point in time it is often expressed as a percentage.

Protease An enzyme that hydrolyzes or cleaves the polyproteins into proteins and is important in the final steps of HIV maturation. In HIV, the protease enzyme is encoded by the *polymerase* gene.

Protease inhibitors A class of ART drugs that interferes with the viral protease enzyme of HIV by inhibiting the viral polyproteins from being cleaved, which would allow the individual viral proteins to produce infectious viral particles.

Reverse transcriptase An enzyme found in HIV that creates double-stranded DNA using viral RNA as a template and host tRNA as primers. The reverse transcriptase enzyme is encoded by the *polymerase* gene of HIV.

Reverse transcriptase (RT) inhibitors A class of ART drugs that interfere with the reverse transcription step during the HIV life cycle. During this step, the HIV enzyme RT converts HIV RNA to HIV DNA. There are two main classes of RT inhibitors that are used as ART drugs.

Nucleoside/nucleotide RT inhibitors (NRTI) are faulty DNA building blocks. When these faulty pieces are incorporated into the HIV DNA (during the process when HIV RNA is converted to HIV DNA), the DNA chain cannot be completed, thereby blocking HIV from replicating in a cell.

Nonnucleoside RT inhibitors (NNRTI) bind to RT, interfering with its ability to convert the HIV RNA into HIV DNA.

RNA (ribonucleic acid) A universal form of genetic material typically transcribed from DNA, it differs from DNA in that it contains ribose and uracil as structural components. In retroviruses like HIV, RNA is their primary genetic material and is found in a mature virus particle.

T-lymphotropic A characteristic of a virus that infects and replicates in T lymphocytes, a type of immune cell. This was the descriptor of the human T cell leukemia virus (HTLV), a human retrovirus that causes T cell leukemia and lymphoma and is T-lymphotropic like HIV. HIV was originally called human T-lymphotropic virus type III (HTLV-III) by Gallo and colleagues.

Tuberculosis The infectious disease caused by *Mycobacterium tuberculosis*. It usually involves the lungs (pulmonary tuberculosis) but can also affect other organs (i.e., kidneys, central nervous system, lymph nodes, bones, etc.; extrapulmonary tuberculosis). Pulmonary tuberculosis, which is the most frequent clinical form, can be classified as smear positive or smear negative according to the result of the sputum bacteriological examination. The former is a major public health problem being highly contagious. Only a few individuals develop tuberculosis after a mycobacterial infection, and most of them soon after infection: it is estimated that the lifetime risk is 5–10% in HIV negatives and 5–15% yearly in HIV positives.

Virion A single and complete extracellular infective form of a virus that consists of an RNA or DNA core and in the case of HIV with a glycoprotein coat or "envelope."

Definition of the Problem

The HIV/AIDS epidemic is now in its third decade since the discovery of the virus responsible for the disease in 1981. While the first cases of AIDS were first recognized in young men who have sex with men in the United States and Europe in the 1980s, it soon became clear that the virus could be spread through contaminated blood products and heterosexual sex. At the time of its discovery, acquired immunodeficiency syndrome (AIDS) was a new disease with high mortality, and the discovery of a new human virus as its cause in 1983 created new challenges for prevention, treatment, and vaccine efforts, many of which remain unmet today. Human immunodeficiency virus type 1 (HIV-1), as the causative agent of AIDS, has been the subject of intense research over the past three decades, in an effort to understand the biological properties of this new virus, its relatedness to other known retroviruses, characterize its epidemiology, and discover drugs and vaccines to control the epidemic.

In the early 1980s, it was recognized that HIV could be spread through blood, blood products, and sexual and perinatal transmission routes. High rates of HIV infection and its accompanying disease began to be recognized throughout the world in the late 1980s as the HIV/AIDS global pandemic became a frightening reality to the international community. The disproportionate burden of infection and disease first recognized in sub-Saharan Africa and then Asia with growing rates of disease and mortality led to estimates of global infections growing in the tens of millions with no magic bullet to end the spread. While global estimates are necessarily fraught with numerous assumptions and subjected to poor reporting, UNAIDS revised down their global estimates in 2007 [1]. Nonetheless, at the end of 2009, an estimated 33.3 million people worldwide were living with HIV; 2.5 million of these were children, with almost two thirds of them in low- and middle-income countries [2]. While stunning in magnitude, these estimates now reflect a continued slowing of the growth trajectory of the pandemic, with lower rates of new infections and mortality, the latter being largely due to the continued scale-up of provision of antiretroviral therapy (ART) that began in 2004.

We are therefore at an important point in the time for the global HIV pandemic; the peak of infections in most countries occurred about a decade ago, and many high-burden countries are seeing a plateau or decrease in HIV prevalence rates. Large-scale treatment programs for low-income countries have succeeded in providing complex and relatively expensive ART drugs to almost one third of patients in need of these life-saving therapies; 5 million of the 15 million in need. Recent strides in identifying new and efficacious methods for prevention have been encouraging as the development of an effective vaccine is still awaited.

Introduction

AIDS was first recognized as a new and distinct clinical entity in 1981 [3–5]. The first cases were recognized because of an unusual clustering of diseases such as Kaposi's sarcoma and *Pneumocystis carinii* pneumonia in young homosexual men. Although such syndromes were occasionally observed in distinct subgroups of the population – such as older men of Mediterranean origin in the case of Kaposi's sarcoma or severely immunosuppressed cancer patients in the case of *Pneumocystis carinii* pneumonia – the occurrence of these diseases in previously healthy young people was unprecedented. Since most of the first cases of this newly defined clinical syndrome involved homosexual men, lifestyle practices were first implicated as the cause of the disease.

AIDS cases were soon reported in other populations as well, including intravenous (IV) drug users [6] and hemophiliacs [7–9]. Hemophiliacs used clotting factor preparations which were prepared from the pooled blood of a huge number of donors, and IV drug users often used needles contaminated with small amounts of blood from previous users, thereby increasing their exposure to foreign tissue antigens. Asymptomatic hemophiliacs and intravenous drug users were often found to have abnormally low CD4 helper lymphocytes and higher than normal T suppressor cells, similar to the gay men with AIDS. The increase in T suppressor cells was presumably due to frequent antigenic stimulation; the decrease in CD4+ T helper cells was the more direct effect of the yet-to-be-discovered causative agent.

Three new categories of AIDS patients were soon observed: blood transfusion recipients [10, 11], adults from Central Africa [12–14], and infants born to mothers who themselves had AIDS or were IV drug users [15, 16]. The transfusion-associated cases had received blood donated from an AIDS patient at least 3 years before they began showing symptoms [10, 11].

Based on the disparate populations afflicted with this new malady and the emerging epidemiology of the disease, the possible infectious etiology for AIDS was considered [17]. Multiple studies were initiated to determine the possible role of various microorganisms, especially viruses in causing AIDS. These studies measured and compared seroprevalence rates for suspect viruses in AIDS patients and controls. The short list of candidate viruses included cytomegalovirus (CMV), because it was already associated with immunosuppression in kidney transplant patients; Epstein-Barr virus (EBV), presumably because it was a lymphotropic virus; and hepatitis B (HBV), because infection with this virus was known to occur at elevated rates in both homosexual men and recipients of blood or blood products. However, based on the unique clinical syndrome and unusual epidemiology of AIDS, if the etiology was an already known virus, it would presumably have to be a newly mutated or recombinant genetic variant.

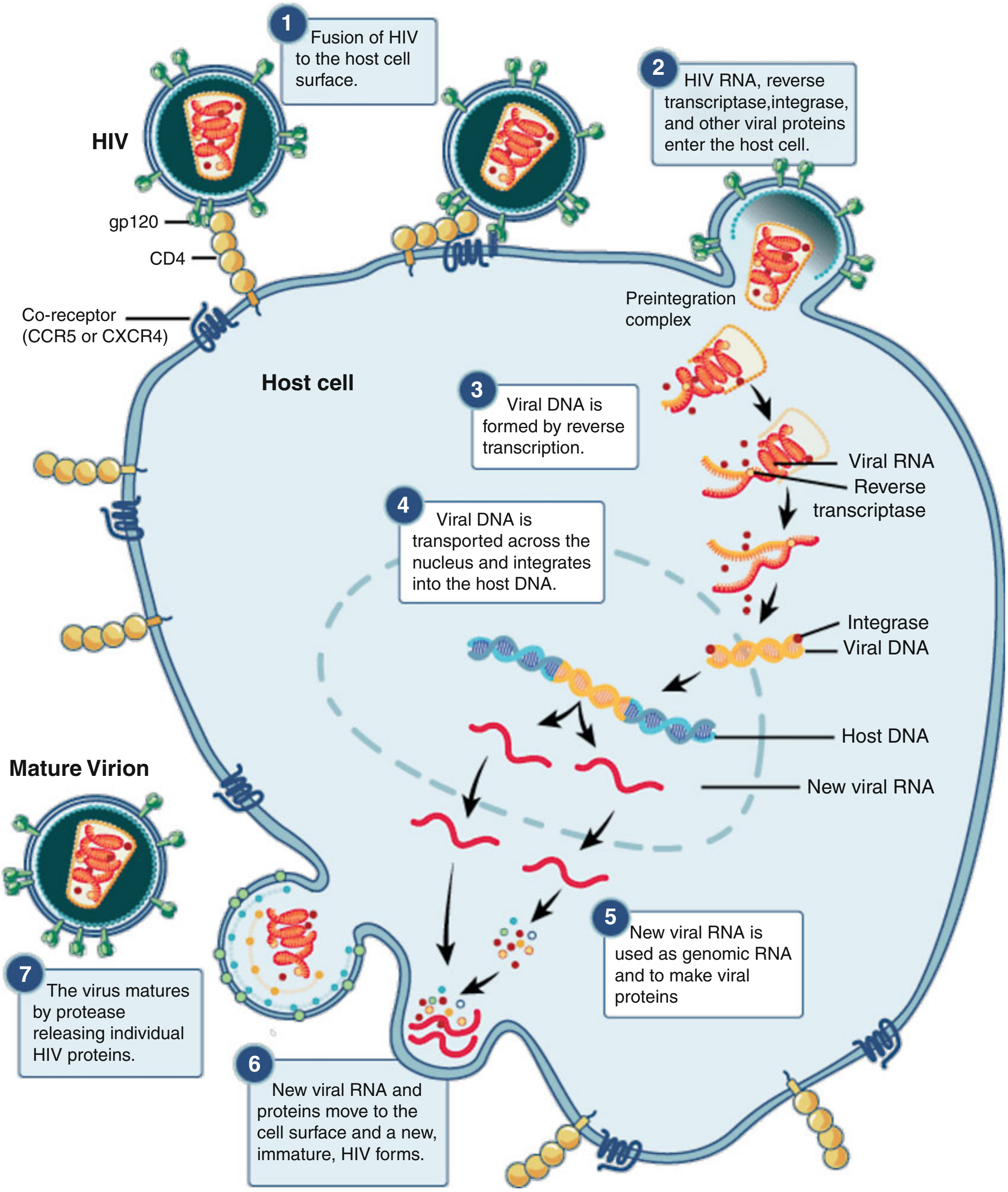
Max Essex [18, 19], Bob Gallo [20, 21], and Luc Montagnier, Francoise Barre-Sinoussi, and Jean-Claude Chermain [22] postulated that a variant T-lymphotropic retrovirus (HTLV) might be the etiologic agent of AIDS. Among the most compelling reasons for this hypothesis was that the human T-lymphotropic retrovirus (HTLV), discovered by Gallo and his colleagues [23] in 1980, was the only human virus known to infect T helper lymphocytes at that time. This fit with the new disease where T helper lymphocytes were selectively depleted by the causative agent [24–26]. AIDS patient blood samples were repeatedly cultured in an attempt to find a virus related to HTLV-I or HTLV-II [27]; however, these studies were only partially successful. Although antibodies cross-reactive with HTLV-I and HTLV-related genomic sequences were found in a minority of AIDS patients [18, 21, 22, 28], the reactivity was weak, suggesting either the coinfection of AIDS patients with an HTLV, or that a distant, weakly reactive virus was the causative agent. Proof that the disease was linked to a T-lymphotropic retrovirus was obtained by Gallo and his colleagues [29–31]. Further characterization of the agent – now termed human immunodeficiency virus type 1 (HIV-1) – revealed that it was the same as the isolate detected earlier by Montagnier and his colleagues [22]. Despite controversy over the names and identity of certain isolates, it is now clear that this new and unique human pathogen was not only a distant genetic relative of the known HTLV but also a virus that may have been more recently introduced into the humans from a primate reservoir.

A second HIV was discovered in 1984 based on antibodies from West African commercial sex workers that recognized proteins not only from the simian immunodeficiency virus (SIV) but also from HIV-1 [32, 33]. It is now known that HIV-1 is more closely related to SIVcpz, found mainly in the *Pan troglodytes troglodytes* chimpanzee species [34], while HIV-2 is related to SIVsm found in sooty mangabey monkeys (*Cercocebus atys*) [35]. HIV-2 is the second human immunodeficiency virus and constitutes the closest known human virus related to the prototype AIDS virus, HIV-1. HIV-2 shares many virologic and biologic features with HIV-1; however, its ability to transmit and cause disease is much lower [36–38]. HIV-2 infection is much less prevalent in the world; it is found primarily in West Africa and other parts of the globe with connections to West Africa, such as India, parts of South America, and urban centers with high rates of immigration from West Africa [39, 40].

Basic Virology of HIV

Human immunodeficiency virus types 1 and 2 (HIV-1 and HIV-2) are members of the *Lentivirus* genus of the

Retroviridae family of RNA viruses and are 50% similar at the genetic level. Retroviruses are enveloped viruses that infect a wide range of vertebrate hosts in a species-specific manner [41]. Steps in the HIV replication cycle (Fig. 1) include (1) recognition and binding of the virion to the host cell surface via specific primary and coreceptors and fusion to the host cell membrane; (2) uncoating of the virion and release of the HIV genetic material and other viral proteins including enzymes such as reverse transcriptase, integrase, and


1
Fusion of HIV to the host cell surface.
2
HIV RNA, reverse transcriptase,integrase, and other viral proteins enter the host cell.
HIV
gp120
CD4
Co-receptor (CCR5 or CXCR4)
Preintegration complex
Host cell
3
Viral DNA is formed by reverse transcription.
Viral RNA
Reverse transcriptase
4
Viral DNA is transported across the nucleus and integrates into the host DNA.
Integrase
Viral DNA
Host DNA
New viral RNA
Mature Virion
5
New viral RNA is used as genomic RNA and to make viral proteins
7
The virus matures by protease releasing individual HIV proteins.
6
New viral RNA and proteins move to the cell surface and a new, immature, HIV forms.

HIV/AIDS Global Epidemic. Figure 1
HIV replication cycle [43]

protease; (3) HIV RNA is reverse transcribed by the enzyme reverse transcriptase to make a double-stranded DNA copy; (4) HIV viral DNA is transported to the cell nucleus, where it integrates into the host cell's DNA, using the HIV viral integrase; (5) a new copy of the HIV viral RNA is produced which will become the genetic material for new HIV virions; (6) new viral RNA and proteins produced by the host cell move to the cell surface, and new, immature virions are packaged; and (7) the viral envelope proteins are inserted into the host cell membrane, and the virions bud from the cell surface, encapsulated by viral envelope, forming a mature HIV virion [42, 43].

Like other retroviruses, HIV's genetic material becomes permanently integrated in the host cell's DNA, resulting in lifelong infection. The enzyme responsible for viral replication and encoded by the *polymerase* gene is the reverse transcriptase, which as an enzyme is error-prone, which results in considerable genetic variation. HIV enters susceptible cells via a primary receptor, the CD4+ on immune cells such as T lymphocytes. However, secondary coreceptors are also required for HIV's entry into a cell [44–47]. These are normal cellular membrane proteins, part of a seven-transmembrane spanning protein family, called chemokine receptors, and involved in the recruitment of chemokines for normal cellular function. CCR-5 and CXCR-4 are chemokine receptors that also serve as major coreceptors for the HIV virus. In conjunction with the primary receptor, CD4 molecule, they allow the virion to fuse with the susceptible host cell and allow virion entry. The CCR-5 coreceptor along with CD4 allows for HIV to infect CCR-5 bearing cells usually of the macrophage/monocyte lineage. The CXCR-4 coreceptor along with CD4 allows for HIV's fusion and entry into T cells [44, 48, 49].

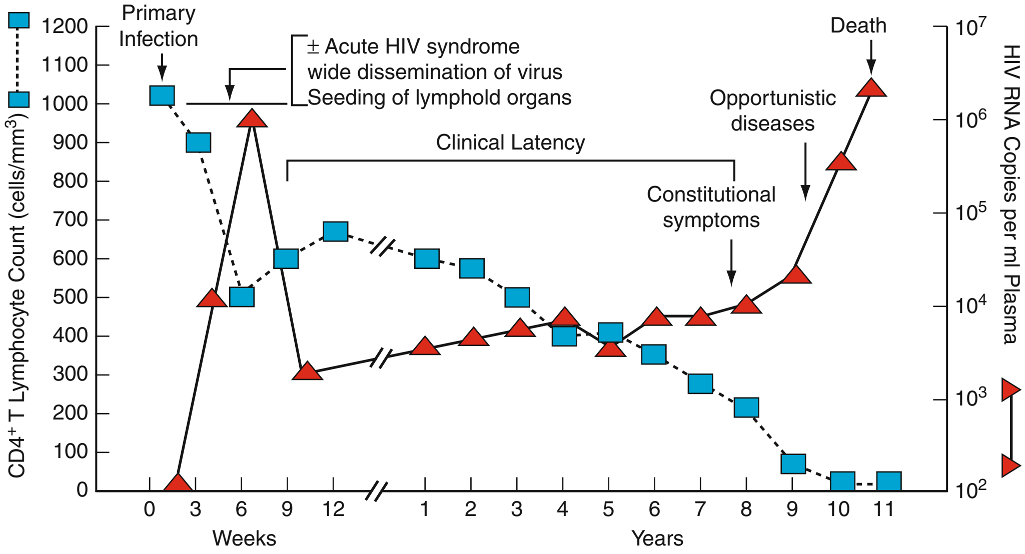
The HIV genome, or genetic makeup, has been characterized and is 9.8 kilobases in length and contains nine different genes, which encode 15 proteins. There are three major classes of proteins: structural, regulatory, and accessory [42]. Three genes, *gag*, *pol*, and *env*, each encode multiple viral structural proteins. The *gag* (group-specific antigen) gene produces a 55-kd precursor protein, which is cleaved during viral maturation into the matrix (MA; p17), capsid (CA; p24), nucleocapsid (NC; p9), p1, p7, and p6 proteins. The *pol* gene (polymerase) encodes the enzymes including protease, reverse transcriptase (RT), and integrase, while the *env* gene (envelope) gives rise to the envelope proteins gp120 and gp41. The regulatory proteins are Tat and Rev, while the accessory proteins include Vpu, Vpr, Vif, and Nef [42].

HIV and the Cause of AIDS

HIV typically enters the body by blood, blood products, or through fluids exchanged in sex or childbirth. The virus infects a large number of CD4+ cells and replicates rapidly. During the acute phase of HIV infection, the blood carries a large number of HIV virus particles, which spread throughout the body, infecting various organs, particularly the lymphoid organs such as the thymus, spleen, and lymph nodes [50, 51]. During this phase, the virus may integrate and hide in the host cell's genetic material. Thus, evading the host's immune system, the virus may remain dormant for an extended period of time. In the acute phase of infection, a significant proportion of people may suffer flu-like symptoms [51, 52] (Fig. 2).

Weeks after exposure to the virus, the immune system responds with killer T cells (CD8+ T cells) and B cell–produced antibodies. At the same time, CD4+ T cell counts rebound. The virus levels may decrease at this point in time perhaps in response to the first host immune response. During this latency phase, a person infected with HIV may not experience HIV-related symptoms for several years despite the fact that the HIV continues to replicate [53]. It is this long course of dormancy that has given the lentiviruses ("lenti" – meaning slow) their name.

The immune system eventually deteriorates to the point that the human body is unable to fight off other infections; the timing and reasons for this are still subject to research. The level of HIV virus in the blood may dramatically increase, while the number of CD4+ T cells drops to dangerously low levels. An HIV-infected person is diagnosed with AIDS when he or she has one or more opportunistic infections, such as



HIV/AIDS Global Epidemic. Figure 2

Clinical progression of HIV infection [52]

tuberculosis, and CD4+ T cells drop below 200 cells per cubic millimeter of blood, where the normal would be ∼1,000 CD4+ T cells.

Epidemiology of the Global HIV Pandemic

HIV and AIDS is a global pandemic. Although the means by which the virus can spread between people remains unchanged, the populations that are highly infected in different parts of the world are different. Transmission of the virus by blood and blood products for the treatment of various diseases such as hemophilia was recognized in the early 1980s. Ensuring that blood banks and blood products were free of HIV was accomplished relatively early in the pandemic in high-income countries, but is yet to be perfected in many low-income countries where screening of HIV and the high burden of disease pose distinct challenges to maintaining safe blood banks and blood products.

Spread of HIV infection through intravenous drug use was also recognized in high-income countries early in the epidemic. These often disenfranchised and stigmatized populations were difficult to identify to provide necessary diagnosis, care, and treatment. This continues to be a challenge for prevention efforts throughout the world, most significant in parts of Asia, where the spread of HIV through this mode of transmission is both prevalent and difficult to control.

Sexual transmission of HIV in most of the high-income countries remains a threat to men who have sex with men, where the virus and its disease were first recognized. In contrast, in most middle- and low-income countries, the risk of HIV transmission is predominantly through heterosexual sex. This means that the risk to women and their offspring is more significant in these geographic locations and represents a distinction in the epidemiology of HIV/AIDS in these settings. As a consequence, the risk of mother-to-child transmission in middle- and low- income countries was most severe in nations already burdened with high infant mortality. In 2009, 370,000 (230,000–510,000) children were infected with HIV through mother-to-child transmission [2]. This represents a decrease of 24% from 5 years earlier due to better methods of preventing transmission through identification of infected pregnant women and provision of more effective prophylaxis. Inadequate access to antenatal and postnatal services remains a barrier in providing these methods to all those in need [54].

UNAIDS has just completed their 2010 report which reports on HIV and AIDS statistics from around the globe ending in 2009 [2] (Table 1). Extensive

HIV/AIDS Global Epidemic. Table 1 UNAIDS regional estimates for adults and children living with HIV (Adapted from [2])

Geographic region	People living with HIV	
	2009	2001
Sub-Saharan Africa	22,500,000	20,300,000
Middle East and North Africa	460,000	180,000
South and Southeast Asia	4,100,000	3,800,000
East Asia	770,000	350,000
Oceania	57,000	29,000
Central and South America	1,400,000	1,100,000
Caribbean	240,000	240,000
Eastern Europe and Central Asia	1,400,000	760,000
Western and Central Europe	820,000	630,000
North America	1,500,000	1,200,000

analysis of these current rates and the trends from previous years allows for some cautious optimism on the status of the pandemic. This report continues to support the notion that the peak of new infections occurred in 1999 and rates of new infections continue to decline or plateau [55, 56]. In 2009, an estimated 2.6 million (2.3–2.8 million) people were newly infected with HIV. In 33 countries, most in sub-Saharan Africa, HIV incidence showed significant decreases compared to 2001 [2].

Globally, an estimated 33 million people are infected with HIV. In earlier years treatment was only provided in high-income countries; more recently local governments and international programs supporting antiretroviral therapy in low-income countries have had an impact on the annual death toll. A cumulative total of 24 million people have died from AIDS between 1980 and 2007, and Bongaarts and colleagues project that this will reach 75 million by 2030 [56].

Regional HIV statistics for both 2009 and 2001 are provided in Table 1. These demonstrate the heterogeneity of the global pandemic. It is beyond the scope of this chapter to discuss each of the region's HIV epidemiology in depth. Summarized findings are therefore provided with more emphasis on sub-Saharan Africa where the highest burden of infection and disease is found.

Sub-Saharan Africa

Over 22 million people in sub-Saharan Africa are infected with HIV, although there is considerable variability among nations in rates of infection and disease [2] (Fig. 3). The countries of Southern Africa (Angola, Botswana, Lesotho, Malawi, Mozambique, Namibia, South Africa, Swaziland, Zambia, and Zimbabwe) have over 11 million people currently infected. Thus, this region of the continent alone represents over one third of the global pandemic in terms of adult HIV infections, new HIV infections, and deaths due to HIV. Swaziland, in 2009, reported a prevalence of 25.9%, representing the highest adult HIV prevalence rate for a given country [57]. In South Africa, where over five million people are living with HIV in 2009, AIDS is the major cause of maternal mortality and attributed to cause over one third of deaths in children under the age of 5 years old.

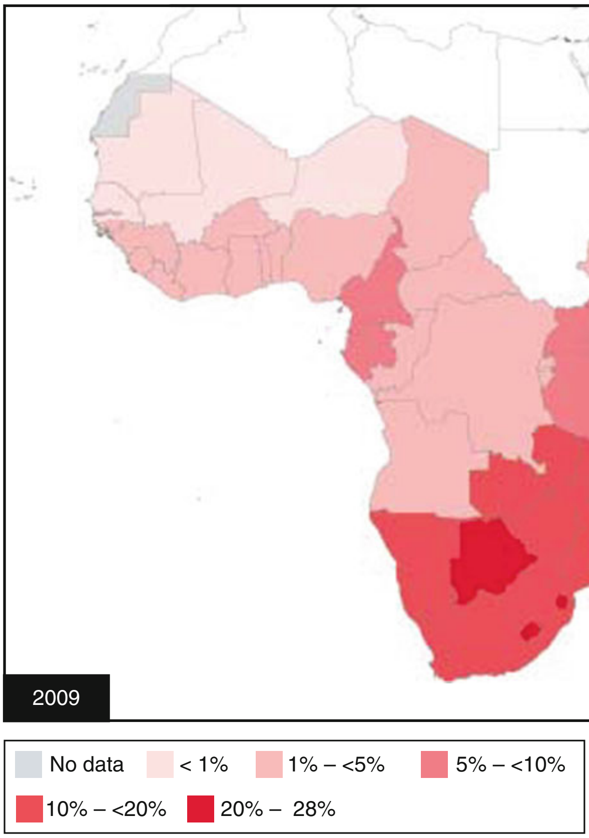
Other large countries, such as Nigeria with a population estimated at 150 million and Ethiopia with a population 82 million, have relatively low HIV prevalence rates below 4%, but still represent a significant portion of the global burden of disease because of their large populations. By contrast, many of the countries of Southern Africa are small, but with high prevalence rates. Thus, the distinct impact of the epidemics in such diverse settings can be appreciated.

As of 2009, the rates of HIV infection in the countries of East Africa have been declining. Uganda's HIV prevalence has been stable in the 6–7% range since 2001. Kenya and Tanzania have shown declining prevalence rates, and Rwanda has been stable at 3% since 2005 [2].

In West and Central Africa, countries have maintained some of the lowest HIV infection rates of the continent. In 2009, 12 countries had rates below 2%; these included: Benin, Burkina Faso, Democratic Republic of the Congo, Gambia, Ghana, Guinea, Liberia, Mali, Mauritania, Niger, Senegal, and Sierra Leone. Nigeria's HIV infection rate has been declining for the past 6 years with the 2009 estimate at 3.6% (3.3–4.0%). The prevalence of HIV is highest in Cameroon at 5.3% (4.9–5.8%); Central African Republic, 4.7% (4.2–5.2%); Côte d'Ivoire, 3.4% (3.1–3.9%); and Gabon, 5.2% (4.2–6.2%) [2].

Asia

HIV infection rates in 2009 for the region are similar to 2004, with an estimated 4.9 million (4.5–5.5 million)



HIV/AIDS Global Epidemic. Figure 3
HIV prevalence rates among adults aged 15–49 years old in sub-Saharan Africa, 2009 (Adapted from [2])

people are currently infected [2]. There is considerable heterogeneity of infection by countries, and some previously low-prevalence countries like Pakistan and Bangladesh are increasing, largely due to transmission in drug-injecting populations. Within countries, there is also considerable variation; for instance, five provinces of a total of 22 in China account for over half of the HIV infections of the country [58, 59]. The epidemic in this region of the world is largely due to significant subpopulations of injecting drug users and men who have sex with men, where rates remain high and often increasing.

Eastern Europe and Central Asia

The largest increases in new HIV infection rates are seen in this region of the world with a burden of 1.4 million people estimated to be living with the virus in 2009 [2]. HIV prevalence exceeds 1% in the Russian Federation and Ukraine. The epidemic is concentrated in high-risk populations that use injection drugs and are involved in sex work and their sexual partners [60].

Caribbean

This region of the world has a relatively low burden of people living with HIV (240,000), yet because of their small population size, their HIV prevalence of ∼1% (0.9–1.1%) is still considered high for the region [2]. There is considerable variability between countries and within countries. Haiti has highly variable rates, including 12% in pregnant women from some urban settings [61, 62], whereas Cuba's prevalence is exceptionally low, at 0.1% (0.08–0.13%) [2]. Unprotected paid sex is considered the major mode of HIV transmission, and infection rates in women are over half of those infected with the virus.

Central and South America

The HIV epidemic appears stable in 2009 compared to previous years, and the number of children infected is declining with ~4,000 children newly infected in 2009 [2]. There are 1.4 million people living with the infection in this region, with concentrated infection rates in men who have sex with men and sex workers.

North America and Western and Central Europe

The burden of HIV infection is 2.3 million (2.0–2.7 million) in these higher-income countries, representing a 30% increase from 2001 [2]. It is believed that unprotected sex among men having sex with men is responsible for these disturbing increasing trends in Canada, the USA, and parts of Europe [63]. General decreases are seen in injecting drug user subpopulations overall, but certain countries continue to see this as a driver of the epidemic [64–66]. Immigrant populations and individuals acquiring HIV from more endemic regions are also contributing to the burden of infection particularly in urban centers [67].

Middle East and North Africa

This region of the world has historically had low rates of HIV infection, which may be due to inadequate reporting. In 2009, an estimated 460,000 people are living with HIV, which is a substantial increase from the 2001 reported statistics [2]. Transmission of the virus from contaminated drug-injection equipment is considered the major mode of transmission in the countries of this region.

Oceania

The burden of HIV infection in this region is small at 57,000 (50,000–64,000) but has increased from the 2001 statistics [2]. Papua New Guinea has the largest infection rates in this region, with a national adult HIV prevalence of 0.9% (0.8–1.0%). Unprotected heterosexual sex is more predominant in Papua New Guinea, whereas unprotected sex in men who have sex with men is more common in New Zealand and Australia [68].

Prevention

The optimal biomedical means of preventing infection with an infectious disease pathogen would be the use of an effective vaccine. As efforts to develop such a vaccine are still underway, public health officials around the world have resorted to other methods of preventing HIV transmission [69]. These revolve around the various types of HIV transmission such as men who have sex with men, sex workers, intravenous drug users, and heterosexual adults in places such as Africa. Education on risks of transmission, methods of prevention, and behavior change are at the core of all prevention programs spanning on populations at risk. For many of these subpopulations, stigma, lack of access to services, or even legislation can diminish the potential of prevention programs and messages for behavior change.

Access to HIV testing and counseling is a foundation to all prevention methods, which provides access to comprehensive HIV care. Education and behavior change are critical components of prevention methods, and provision is often most effective when provided by community groups. In the early 1980s in the USA, pioneer groups such as the San Francisco AIDS Foundation, Gay Men's Crisis in New York, and AIDS Project Los Angeles demonstrated the role of education in promoting safer sex practices among gay and bisexual men [70]. Recent increases in new infections in young gay men in North America and Europe are of concern and suggest the need to enhance prevention efforts [2, 63].

Globally, HIV acquired through injecting drug use is estimated to represent 20% of the people living with HIV in 2009 [2]. Efforts to make injecting drug use safer through provision of clean needles have proven efficacy in lowering transmission through this route. UNAIDS reports in 2009 that in Eastern Europe and Central Asia where injecting drug use is a major driver of the HIV epidemic, five of nine countries report that more than 80% of people injecting drugs used sterile injecting equipment at their last injection. Eight of twelve countries in South and Southeast Asia reported the same [2]. However, the various governmental responses across the globe have often prevented the widespread adoption and support of such programs [65]. While provision of clean needles is inexpensive, it is believed that many programs provide inadequate numbers of clean needles to insure optimal prevention [66]. In addition, the substance abuse treatment programs, such as methadone programs, may help curtail the intravenous drug–using behavior and therefore eliminate the risk of the transmission [64].

Prevention methods directed at commercial sex work are needed worldwide and again represent a targeted high-risk population that is disenfranchised and difficult to reach through conventional health system structures. The prevention methods focus on HIV counseling and testing, safe sex through proper condom use, and concomitant diagnosis and treatment of sexually transmitted infections (STIs), which are known to facilitate HIV transmission [71]. A longstanding program for sex workers in Southern India, called Avahan, provided a combined prevention approach of community outreach, empowerment, condom programming, and STI and HIV testing services over an 8-year period with a decrease in HIV prevalence from 20% to 16% in these high-risk women [71, 72].

In many parts of the resource-limited countries, heterosexual transmission is the major risk factor for HIV acquisition. Testing and access to HIV care is again critical to these populations along with appropriate counseling on behavior change, proper condom use, and STI diagnosis and treatment. Proven prevention measures for this group would include abstinence, mutual monogamy, reduced number of sexual partners, and consistent condom use. For young people, abstinence and older age of sexual debut are considered part of prevention messages [69].

Since 2000, a number of epidemiologic surveys and studies considered the possible benefit of male circumcision (surgical removal of the foreskin) in decreasing HIV transmission [73]. The penile foreskin is thought to be sensitive to epithelial tears, which carry a high density of CD4-bearing cells as targets for virus infection and subsequent replication [74]. In addition, uncircumcised men are at higher risk for STIs, which would facilitate HIV transmission [75–77]. In 2005–2006, randomized clinical trials conducted in South Africa, Kenya, and Uganda demonstrated a 55–76% lower HIV incidence compared to men still on the wait list of circumcision [78–80]. It is still unknown whether male circumcision will decrease male-to-female transmission of the virus, but these studies are underway [80]. There are other health benefits to male circumcision, but the surgical procedure is not without some risk, although serious complications are rare [76]. Nonetheless, for populations where men or boys are not circumcised, such as much of Southern and Eastern Africa, this becomes an important prevention tool [81].

Mother-to-child transmission of HIV is in large part responsible for the 2.5 million children living with HIV worldwide. Prevention of mother-to-child transmission (PMTCT) consists of identification of HIV infection in pregnant women and provision of a short course of antiretroviral drugs in a prophylactic manner (rather than treatment of the mother for her own disease to lower the viral burden in the mother and to decrease the risk of transmission to the baby). In resource-rich countries, most HIV-infected pregnant women would receive full ART (three drugs) as a means of preventing mother-to-child transmission; however, in most resource-limited settings, mono- or bi-therapy ART has been used due to cost and international recommendations at the time [82, 83]. In 2010, WHO revised their PMTCT guidelines based on clinical trial data from Africa, indicating that full ART could virtually eliminate mother-to-child transmission [84]. In settings where breast-feeding is judged to be the safest infant feeding option, ART prophylaxis continued to the mother and baby during this period could severely limit this mode of transmission [82–84]. While access to any PMTCT services remains a problem in a number of resource-limited countries and settings, only 54% access for sub-Saharan Africa, the potential for virtual elimination of mother-to-child transmission as these new guidelines are implemented becomes a real possibility [2, 82].

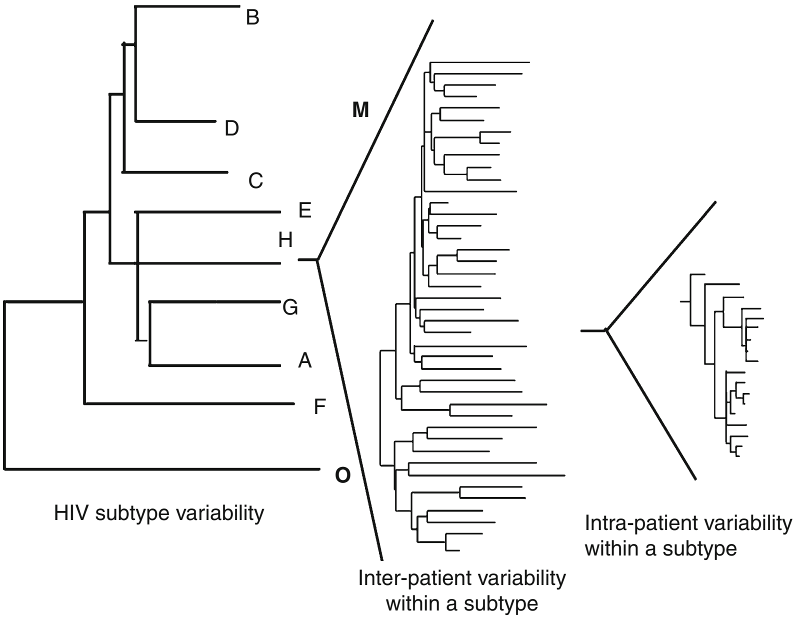
Postexposure prophylaxis (PEP) was the terminology frequently used to describe the use of short-course ART drugs given after an occupational setting exposure such as needle sticks with potentially HIV-infected blood. Similarly, PEP has been used in rape cases in settings where HIV infection is high and the potential for sexual transmission can be potentially decreased [85, 86]. Preexposure prophylaxis (PrEP) is the use of an ART drug administered to an HIV-negative individual prior to or at the time of the potential transmission event in an effort to decrease the viral inoculum and limit the potential for infection.

A recent study reported from South Africa this past year described the use of tenofovir-based gel as a microbicide to be used once 12 h before sex and again 12 h after sex as both safe and effective in reducing HIV transmission and STIs [87–89]. PrEP has also been successful in decreasing transmission in men who have sex with men and in heterosexual discordant couple studies [90, 91]. This paves the way for additional biomedical means of preventing infection, which can be coupled with other prevention methods depending on the setting and populations at risks [92, 93]. Further trials are underway to test this new and promising strategy to prevent HIV transmission.

Genetic Diversity of HIV

Early in the history of HIV research, it was recognized that each viral isolate varied from another in its nucleotide sequence, and multiple mechanisms for this genetic variation were considered [42, 94]. Part of the HIV replication cycle requires the reverse transcription of viral RNA by the virus' reverse transcriptase. This enzyme is error-prone, thereby generating approximately 1 error per 10^4 nucleotides, or viral genome. Estimates of HIV replication in people have suggested 10^{10}–10^{12} new viral particles produced per day [95, 96]. Thus, the potential for viral variation is significant even if the immune system, tissue compartmentalization, or antiretroviral therapy does not select for particular viral variants. Genetic variability between viral isolates from different patients [97, 98], frequently termed interisolate or interpatient variability, can vary by up to 5%. And in turn, this was also distinguishable from the genetic variation that was seen at the level of an individual patient, often termed intrapatient variability and variability up to 1% [99]. At the level of the individual patient, a swarm or quasispecies of highly related but distinguishable viral variants has been demonstrated throughout the course of HIV infection [100–102]. Thus, the genetic variation of HIV is hierarchical as depicted in the simplified schematic of subtype variation, interpatient variation, and intrapatient variation (Fig. 4).

Phylogenetic analyses of viruses from various geographic regions were used to identify three distinct groups M, N, and O within HIV-1 [34, 103, 104]. Most HIV-1 sequences belong to group M (major). A divergent subset of viruses identified in Cameroon in 1994, which did not cluster with group M viruses, were classified as group O (outlier) [105], and in 1998, another set of viruses which did not cluster with group


B
D
C
E
H
G
A
F
O
M
HIV subtype variability
Inter-patient variability within a subtype
Intra-patient variability within a subtype

HIV/AIDS Global Epidemic. Figure 4

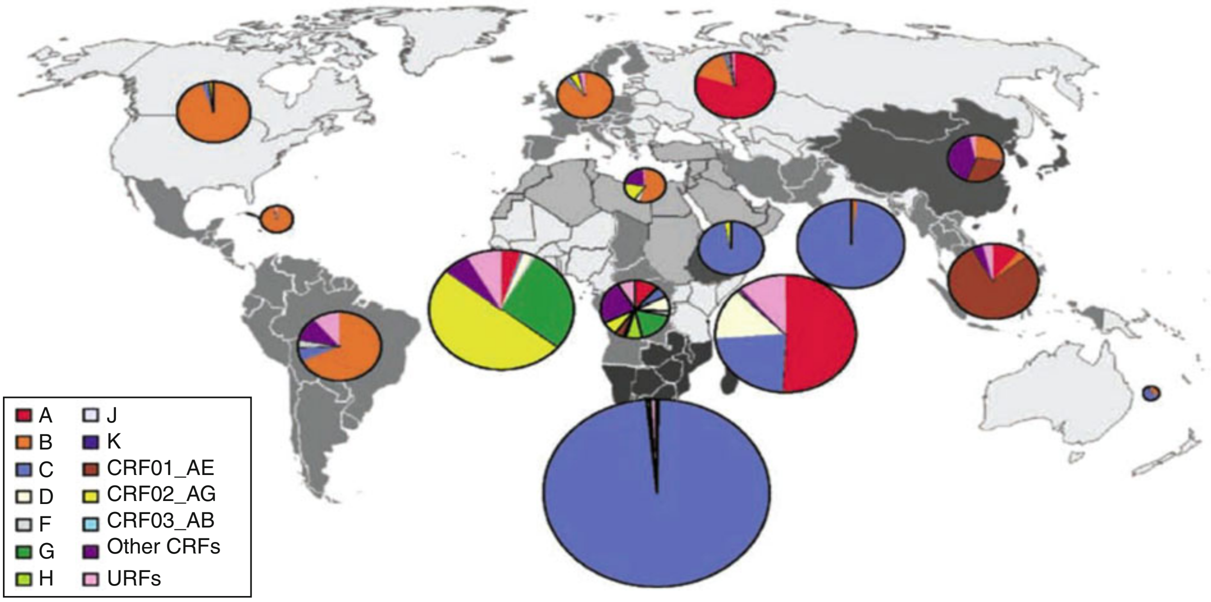
Phylogenetic trees depict differences between HIV subtypes, variation between viruses from different patients, interpatient variability, and variation between viruses within the same patient, intrapatient variability

M or O viruses were termed group N viruses [106]. A recent virus characterized from a gorilla has been designated group P [107]. It is thought that all virus groups were introduced by independent SIVcpz or SIVgor transmissions into the human population in the early part of the twentieth century [34, 104, 108].

The fruits of international HIV research have provided information of HIV viruses from different parts of the globe [109] (Fig. 5). The HIV-1 group M has been subdivided into genetic subtypes, defined originally on comparison of gag and env sequences of the virus and then later with full-length sequencing of the virus [110, 111]. Within a subtype, the average env genetic variation ranges from 5% to 15%, whereas the average env genetic variation between subtypes ranges from 20% to 30% [112, 113]. Remarkably, all the HIV-1s from the USA and Western Europe have been of a single subtype, B. Most of the diverse subtypes of HIV-1 including circulating recombinant forms (CRFs) and unique recombinant forms (URFS) have been found in sub-Saharan Africa and, to a lesser degree, Asia, now accounting for 20% of the worldwide HIV viruses [109].

In Thailand, HIV-1 subtype B was detected in IV drug users during the mid-1980s. During the late 1980s, subtype E was first detected. By the early to mid-1990s, HIV-1 subtype E (CRF01_AE) had spread very rapidly throughout heterosexuals in Thailand, with the highest rates in the northern regions of the country [114]. Although apparently present earlier in the region, HIV-1 subtype B never spread to cause a major heterosexual epidemic, as did HIV-1 subtype E. Full-length sequencing of subtype E determined that it is actually a circulating recombinant form, consisting of subtypes A and E, now referred to as CRF01_AE [115, 116].

A similar situation occurred in India, with HIV-1 subtypes B and C. While B appeared to be introduced earlier and it expanded among intravenous drug users, this subtype did not appear to spread as rapidly among heterosexuals as did HIV-1 C. Previously associated with



HIV/AIDS Global Epidemic. Figure 5

Regional distribution of HIV-1 subtypes and recombinant viruses, 2004–2007. Fifteen regions of the world are depicted with *pie charts* representing the distribution of HIV-1 subtypes and recombinants from 2004 to 2007. The size of the pie charts corresponds to the relative numbers of people living with HIV in the region. *CRF* circulating recombinant form, *URF*, unique recombinant form (Adapted from [109]) the massive heterosexual epidemic in southeastern Africa, subtype C also caused a rapid heterosexual epidemic in Western India, apparently initially spreading from the Bombay region [114, 117]. The results in Africa and Asia suggest that HIV-1 subtypes A, C, D, and CRF01_AE are well adapted for heterosexual transmission compared to subtype B, perhaps due to differential ability to infect dendritic cells in the sexual mucosa [118].

Various studies have shown associations between viral genotype and biology [119]. There is some evidence suggesting a relationship between subtype and modes of transmission. Studies in Cape Town [120], Finland [121], Thailand [122, 123], and Australia [124] found that most subtype B strains were associated with homosexual transmission while non-B strains were associated with heterosexual transmission. Infection with certain subtypes has also been associated with increased risk of vertical transmission. A study conducted on mother-child pairs in Tanzania revealed that mothers infected with HIV-1 subtype A, subtype C, and intersubtype recombinant were more likely to transmit to their infants than mothers infected with subtype D [125]. In an earlier related study, Renjifo et al. [126] also found that in perinatally transmitted C/D recombinant viruses, the V3 regions (*env*) were always from subtype C and never from subtype D, suggesting that viruses containing subtype D-V3 have a reduced fitness as compared to those with subtype C-V3. Finally, a study of injection drug users in Thailand found a significantly higher transmission probability associated with subtype E (CRF_01 AE) as compared to subtype B [127].

Some studies have also demonstrated significant differences between subtypes with regards to disease progression. Kanki et al. [128] found that women infected with a non-A subtype were eight times more likely to develop AIDS than were those infected with subtype A. Similarly, Kaleebu et al. [129, 130] reported that subjects with subtype A had a slower progression to disease than those with subtype D. Clinical and immunological differences have also been found between subtypes. In Kenya, where subtypes A, C, and D were all cocirculating within the same population, Neilson et al. [131] found that high plasma RNA levels and low CD4 counts were significantly associated with subtype C infection. In a prospective study conducted at a methadone treatment clinic in Thailand, people infected with CRF01_AE were found to have higher viral loads in early infection than those infected with subtype B [132]. However, this difference decreased over time such that the viral loads were similar at 12, 18, and 24 months postseroconversion [132]. Similarly, a study in our laboratory indicated that women infected with CRF02_AG had a significantly higher viral load during the early stage of infection than women not infected with CRF02_AG [133]. Infection with multiple subtypes has also been associated with higher viral load and lower CD4+ T cell counts [134].

Complete sequence analysis of viruses from around the world has further characterized an HIV pandemic that is increasing in complexity with a higher prevalence of recombinant viruses, currently at 20% of all HIV viruses [109]. In West Africa, Asia, and South America, the predominant virus is a circulating recombinant form. The predominant virus of West Africa is CRF02_AG, which appears to represent a recombinant event that occurred early in the divergence of subtypes A and G. In contrast, the BC and BF recombinants of China and South America, respectively, are derived from more recent recombinants between contemporary parental virus lineages [135]. An understanding of the genetic diversity of predominant HIV-1 subtypes, sub-subtypes, and circulating recombinant forms in a given population may be important in designing effective HIV vaccines [136, 137]. Although the importance of matching a vaccine candidate to regional circulating strains is yet unclear, incorporation of local strains might maximize the efficacy of a potential vaccine candidate [137].

Treatment

Antiretroviral therapy (ART) drugs that block or reduce HIV virus replication in the infected patient are used for the treatment of AIDS. The first ART drug was put into use in 1986–1987: zidovudine (AZT), a nucleoside reverse transcriptase inhibitor (NRTI) that because of its chemical similarity to the host's nucleoside, inhibited the growing DNA chain and served as a chain terminator such that virus replication could not occur. Other NRTIs act in a similar manner, but another class of ART drugs includes nonnucleoside reverse transcriptase inhibitors (NNRTI), which bind to the virus' reverse transcriptase enzyme in a specific position and manner, blocking the enzyme from completing the transcription. NRTI and NNRTIs are often used in combination as the typical first-line regimens for patients with AIDS [138–140]. Since AZT, there has been tremendous progress in developing improved and more efficacious ART drugs and combination regimens for treating AIDS patients in high-income regions of the world; however, the cost of many of these newer drugs has been prohibitive for most middle- and low-income countries despite their disproportionate share of the AIDS burden.

Currently, there are a total of five classes of FDA-approved ART drugs that act on different parts of the HIV virus life cycle: NRTIs and NNRTIs act on the reverse transcription; other drugs act by inhibiting the viral protease, the viral integrase, or blocking virus entry at the level of the coreceptor. ART is typically prescribed in combination with current recommendations requiring at least two active classes of drugs [138–140]. Side effects and toxicities for each of these drugs may be serious and result in significant complications, requiring the patient to switch drug regimens.

The error-prone reverse transcription of HIV replication and its viral dynamics suggest that every possible single point mutation in the HIV genome occurs between 10^4 and 10^5 times per day in an untreated HIV-1-infected individual [141]. Therefore, the propensity for development of drug resistance mutations is high and the reason that strict adherence to taking the prescribed drugs is necessary to maintain viral suppression. In an effort to improve adherence and decrease pill burden to the patients, fixed-dose combination pills are available for most of the typical first-line ART regimens, allowing all drugs to be included in a single pill.

By the late 1990s, treatment for AIDS was widely available in high-income countries, but the vast majority of patients in middle- and low-income countries could not access these expensive drugs. In 2000, at the United Nations Millennium Summit, Kofi Annan declared a call for action to address the gap in treatment and to promote prevention and research in response to the AIDS epidemic [142]. In 2001, the UN General Assembly adopted a Declaration of Commitment on HIV/AIDS, endorsing equitable access to care and treatment as a fundamental component of a comprehensive and effective global HIV response [143]. UNAIDS had reported at the end of 2002 that developing countries had ~300,000 AIDS patients receiving ART with an unmet need estimated at 5.5 million [144]. The international donor community moved toward addressing the goals of equitable access through initiation of large-scale treatment and care programs. The President's Emergency Plan for AIDS Relief (PEPFAR), a program launched in late 2003, authorized over $15 billion dollars in 15 target countries, the largest commitment to a single disease in the history of US government support [145]. Other programs such as the Global Fund to Fight AIDS, Tuberculosis and Malaria [146], the Bill & Melinda Gates Foundation [147], Clinton Foundation [148], World Bank [149], and Medecins Sans Frontieres [150], to name a few, supported this effort with advocacy to country governments, and the rapid scale-up of providing ART to patients in need moved from rhetoric to action.

UNAIDS reported at the end of 2009 that 5.2 million people in low- and middle-income countries received ART, a significant increase from 2008 and unimaginable in 2003 [2]. However, the pandemic has continued, and in these countries, this accomplishment only represents 36% of the 15 million AIDS patients in these countries in need. Children and patients coinfected with TB remain an ART priority in these settings. Many of the original ART drugs developed in the late 1980s went off patent and have now been manufactured in generic form and meet international standards [151]. The Clinton Foundation in mid-2000 helped negotiate important drug cost reductions so that first-line drug per patient approximated $350 per year, rather than the $1,000 per month of the brand name drugs in 2000 [148]. However, the improvement in ART therapy in high-income countries was made possible through newer drugs and new classes of drugs like integrase inhibitors, or entry blockers, which are still cost-prohibited in most middle- and low-income country ART programs. Thus, although patients are receiving ART, the pharmacy of drugs available for toxicities or second- and third-line therapy remains limited in most low-income countries.

ART is not just the provision of pills; the diagnosis of AIDS requires laboratory tests and the clinical management of an AIDS patient entails regular monitoring of immune function, diagnosis, and treatment of comorbidities and measurement of viral load (virus levels in the bloodstream) when possible [142]. These laboratory tests often require sophisticated lab equipment and costs per test that can be quite high, almost the same cost of annual drug costs. In many low-income settings, these diagnostic or monitoring tests were not even available; in some instances where patients were required to pay for clinical tests, the costs were prohibitive to the typical AIDS patient that already had significant economic burdens. In recent years, a major contribution of large-scale care and treatment support from PEPFAR and others has built the infrastructure and provided the training to perform these critical laboratory tests. Provision of laboratory equipment, insuring adequate clean water and electricity to run the equipment and laboratory test specific requirements for cold chain, and short expiration remain challenges for both start-up and sustainability. In some instances, the costs of the tests have been reduced, but there is a critical need for cheaper, efficient, and point-of-care tests that will increase their availability and use for patient management.

Further, the complexity of clinical management and care of HIV infection requires training, infrastructure building, and overall strengthening of the existing health-care system in most developing country settings. Again, the leadership in individual countries coupled with support from the international donor communities has provided the training necessary for the complex administration of HIV care and treatment. However, human resources in many countries are limiting, and further efforts to encourage task-shifting within the medical team and bolstering of the medical education system are important to achieving the provision of ART to the many patients that are still in need.

Vaccine

Since the discovery of the virus and its initial characterization in the early 1980s, the search for an effective vaccine has been the subject of intense research. HIV infection can occur via free virus and/or virus-infected cells. Once infection has occurred with integration of viral genetic material in the target host cell, elimination of the virus by the host's immune mechanisms is not possible. The approach to vaccine design has often sought to mimic the protective immune responses, often identified in animal model systems or in individuals that have demonstrated protection [152, 153]. To date, the correlates of protective immunity from HIV have yet to be identified. Vaccine scientists therefore began with more conventional approaches of seeking to design vaccine candidates that would elicit strong virus-neutralizing antibodies or robust cell-mediated responses or both [154, 155]. The proof of principle has been shown for both of these approaches in nonhuman primate model [156, 157]. However, the model has not been predictive of HIV vaccine efficacy in humans; differences in routes of exposures, strains, and amount of virus are just some of the complexities of this model. Dozens of vaccine candidates have been developed, tested in nonhuman primate models and even early clinical trials in humans [157, 158].

AIDS vaccine candidates are evaluated in a stepwise manner in a series of clinical trials known as phases I, II, and III. Phase I and II trials generally involve a small number of volunteers and provide researchers with critical information about the safety and immunogenicity of the vaccine. The cost of phase III trials can be in hundreds of millions of dollars, and the enrollment for the trial is large, so the AIDS vaccine community moved to an intermediate "test of concept" trial where lower numbers of trial participants were needed; the statistical power of the trial design would allow evaluation of immunogenicity and low ranges of efficacy. It is not until phase III trials that the efficacy of the vaccine is truly assessed.

Candidate HIV vaccines that were tested in phase I and II trials over the past 20 years included inactivated whole virus, or virus particles, subunit HIV envelope vaccines delivered in novel adjuvant or delivery systems, DNA vaccines and live recombinant vector vaccines, and various combination approaches. The results of phase I and II trials indicated that they were safe but did not generate sufficient immune responses to warrant proceeding to phase III trials [157, 159–163].

By 2008, only four candidate HIV vaccines had proceeded to phase IIb and III efficacy trials. Two of the trials were subunit gp120 (envelope) vaccines tested by VaxGen in the USA, Canada, and the Netherlands with 5,400 volunteers, largely men who have sex with men [164]. A second trial was conducted in 2,545 intravenous drug users in Thailand, using recombinant envelope from subtype B and also CRF01_AE [165, 166]. Although neutralizing antibodies were induced by the vaccine, neither trial showed evidence of protection. Subsequently, two trials assessed vaccine candidates based on the premise that cellular immune responses would be protective; the viral antigens were delivered in an adenovirus 5 vector; in order to enhance the generation of the cellular responses, the candidates were tested in the "Step" and "Phambili" trials conducted by Merck [167–169]. Adenovirus 5 is considered a harmless virus, but it does infect humans, and many populations mount antibodies to this virus, frequently without any signs of disease. Some of the volunteers in the Step and Phambili trials did not have antibodies to adenovirus 5, and in those volunteers, the vaccine showed no effect. However, in volunteers with antibodies to adenovirus 5, the vaccinated group showed higher rates of HIV infection, and thus, the vaccine had increased rather than decreased the chances of HIV infection [167–170]. Both trials were interrupted as a result of these unexpected results.

In 2008, a Summit on HIV Vaccine Research and Development determined that research toward a vaccine needed to place greater emphasis on basic research and slowed the pace of candidates, moving toward expensive clinical trials in people [171, 172].

In late 2009, the results of the US military vaccine group's RV144 trial in Thailand were announced. A phase IIb trial was conducted with a recombinant vector approach using the ALVAC virus (an avian poxvirus) used to express HIV proteins to prime the immune response combined with recombinant gp120 protein; both of these entities alone had been tested in people and shown poor immunogenicity [173]. The RV144 trial conducted in 16,400 men and women in Thailand starting in 2003, nonetheless, showed a modest efficacy of 31% protection after 3.5 years of follow-up; there were 51 infected in the vaccine group compared with 74 in similarly sized placebo group [174]. Further analysis of those protected failed to show significant neutralizing antibodies or even CD8+ T cell responses although more had favorable CD4+ T cell responses. Thus, the trial, while giving some optimism for vaccines, has failed to clearly indicate what specific type of immunity was responsible for the protection seen; further analysis will continue to evaluate the trial in hopes of guiding more improved candidates for the future [175].

While the search for a vaccine continues, the research thus far has given not only the scientific community a deeper understanding of both the HIV virus but also the human immune system and its interaction with this unique virus [157]. This will not only inform vaccine efforts for other infectious diseases but also facilitate the use of novel immunomodulatory entities and delivery systems that have been developed and tested. The development of an effective HIV vaccine will be used in conjunction with the many established methods of prevention that have been described, and new modalities such as male circumcision and ART treatment as prevention, which show great promise for additional preventative benefits [176]. Advances in these areas have now merged the interventions of treatment, vaccine, and prevention, to form a future strategy of combined prevention [69].

Future Directions

Thirty years ago, a mysterious disease affecting men who have sex with men in the USA initiated the world's appreciation of a new human virus with disease-causing potential and global spread that was unprecedented. Since that time, it is estimated that 60 million people have been infected with HIV responsible for over 25 million deaths. The history of HIV's discovery is instructive to our overall outlook on infectious diseases in general. HIV originated in closely related viruses of primates with multiple variants that still resemble close relationship to the viruses of great apes. The virus and its many genetic variants showed distinct differences in its epidemiology based on geographic location, risk groups affected, spread, and disease. In addition, the increasing complexity and higher prevalence of recombinant forms of the HIV epidemic suggests that more vigilant surveillance with improved technologies will be important. The marked dichotomy between the epidemics of high-income versus middle- and low-income countries has challenged public health efforts to address prevention and treatment needs. ART drugs and laboratory tests developed in high-income countries are high in cost and complexity of use; this continues to be a major obstacle to the largest number of people infected and suffering from HIV and AIDS. In the past decade, an international effort to provide equal access to prevention, care, and treatment has demonstrated that the gap in access can be narrowed. Newer and more efficacious regimens to prevent mother-to-child transmission have just been adopted in the 2010 WHO guidelines and hold the potential of eliminating transmission as the guidelines are rolled out [84]. WHO has also revised the 2010 ART guidelines, which will initiate treatment for patients with less severe disease, CD4+ cell counts below 350, rather than 200 cells per cubic milliliter of blood [177]. While the increased costs and actual implementation of these new guidelines will pose an obstacle for many poor countries of the world, the ultimate realization of these guidelines will further decrease the burden of people living with HIV and increase the longevity of those already on treatment.

The historic and unprecedented support and efforts from the international donor community have made major progress in scaling up prevention and treatment of HIV in areas of the world with the most need. They have begun the difficult task of encouraging national leadership, building up broken or nonexistent health systems, in order to address the pandemic, but the burden of disease is still significant, and efforts must be continued with a perspective of long-term sustainability.

Bibliography

1. UNAIDS (2007) AIDS epidemic update. http://data.unaids.org/pub/epislides/2007/2007_epiupdate_en.pdf
2. UNAIDS (2009) AIDS epidemic update. http://data.unaids.org/pub/Report/2009/JCI700/Epi Update 2009 en.pdf
3. Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA, Saxon A (1981) Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. N Engl J Med 305:1425–1431
4. Masur H, Michelis MA, Greene JB, Onorato I, Stouwe RA, Holzman RS, Brettman L, Lange M, Murray HW, Cunningham-Rundles S (1981) An outbreak of community-acquired *Pneumocystis carinii* pneumonia: initial manifestation of cellular immune dysfunction. New Eng J Med 305:1431–1438
5. Siegal FP, Lopez C, Hammer GS, Brown AE, Kornfeld SJ, Gold J, Hassett J, Hirschman SZ, Cunningham-Rundles C, Adelsberg BR et al (1981) Severe acquired immunodeficiency in male homosexuals, manifested by chronic perianal ulcerative herpes simplex lesions. New Eng J Med 305:1439–1444
6. Anonymous (1982) Epidemiologic aspects of the current outbreak of Kaposi's sarcoma and opportunistic infections. New Eng J Med 306:248–252
7. Davis KC, Horsburgh CR Jr, Hasiba U, Schocket AL, Kirkpatrick CH (1983) Acquired immunodeficiency syndrome in a patient with hemophilia. Ann Int Med 98:284–286
8. Poon MC, Landay A, Prasthofer EF, Stagno S (1983) Acquired immunodeficiency syndrome with *Pneumocystis carinii* pneumonia and Mycobacterium avium-intracellulare infection in a previously healthy patient with classic hemophilia. Clinical, immunologic, and virologic findings. Ann Int Med 98:287–290
9. Elliott JL, Hoppes WL, Platt MS, Thomas JG, Patel IP, Gansar A (1983) The acquired immunodeficiency syndrome and Mycobacterium avium-intracellulare bacteremia in a patient with hemophilia. Ann Int Med 98:290–293
10. Curran JW, Lawrence DN, Jaffe H, Kaplan JE, Zyla LD, Chamberland M, Weinstein R, Lui KJ, Schonberger LB, Spira TJ et al (1984) Acquired immunodeficiency syndrome (AIDS) associated with transfusions. New Eng J Med 310:69–75
11. Jaffe HW, Francis DP, McLane MF, Cabradilla C, Curran JW, Kilbourne BW, Lawrence DN, Haverkos HW, Spira TJ, Dodd RY, Gold J, Armstrong D, Ley A, Groopman J, Mullins JI, Lee TH, Essex M (1984) Transfusion-associated AIDS: serologic evidence of human T-cell leukemia virus infection of donors. Science 223:1309–1312
12. Piot P, Quinn TC, Taelman H, Feinsod FM, Minlangu KB, Wobin O, Mbendi N, Mazebo P, Ndangi K, Stevens W (1984) Acquired immunodeficiency syndrome in a heterosexual population in Zaire. Lancet 2:65–69
13. Van de Perre P, Rouvroy D, Lepage P, Bogaerts J, Kestelyn P, Kayihigi J, Hekker AC, Butzler JP, Clumeck N (1984) Acquired immunodeficiency syndrome in Rwanda. Lancet 2:62–65
14. Clumeck N, Mascart-Lemone F, de Maubeuge J, Brenez D, Marcelis L (1983) Acquired immune deficiency syndrome in Black Africans (letter). Lancet 1:642
15. Oleske J, Minnefor A, Cooper R et al (1983) Immune deficiency syndrome in children. J Am Med Assoc 249:2345–2349
16. Scott GB, Buck BE, Leterman JG et al (1984) Acquired immunodeficiency syndrome in infants. New Eng J Med 310:76–81
17. Francis DP, Curran JW, Essex M (1983) Epidemic acquired immune deficiency syndrome (AIDS): epidemiologic evidence for a transmitted agent. J Natl Cancer Inst 71:1–6
18. Essex M, McLane MF, Lee TH, Falk L, Howe CWS, Mullins J, Cabradilla C, Francis DP (1983) Antibodies to cell membrane antigens associated with human T-cell leukemia virus in patients with AIDS. Science 220:859–862

19. Essex M, McLane MF, Lee TH, Tachibana N, Mullins JI, Kreiss J, Kasper CK, Poon M-C, Landay A, Stein SF, Francis DP, Cabradilla C, Lawrence DN, Evatt BL (1983) Antibodies to human T-cell leukemia virus membrane antigens (HTLV-MA) in hemophiliacs. Science 221:1061–1064
20. Gelmann EP, Franchini G, Manzari V, Wong-Staal F, Gallo RC (1984) Molecular cloning of a unique human T-cell leukemia virus (HTLV-IIMo). Proc Natl Acad Sci USA 81:993–997
21. Gallo RC, Sarin PS, Gelmann EP, Robert-Guroff M, Richardson E, Kalyanaraman VS, Mann D, Sidhu GD, Stahl RE, Zolla-Pazner S, Leibowitch J, Popovic M (1983) Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). Science 220:865–867
22. Barre-Sinoussi F, Chermann J-C, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Vezinet-Brun F, Rouzioux C, Rozenbaum W, Montagnier L (1983) Isolation of T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220:868–870
23. Poiesz BJ, Ruscetti FW, Reitz MS, Kalyanaraman VS, Gallo RC (1981) Isolation of a new type C retrovirus (HTLV) in primary uncultured cells of a patient with Sezary T-cell leukaemia. Nature 294:268–271
24. Ammann AJ, Abrams D, Conant M, Chudwin D, Cowan M, Volberding P, Lewis B, Casavant C (1983) Acquired immune dysfunction in homosexual men: immunologic profiles. Clin Immunol Immunopathol 27:315–325
25. Fahey JL, Prince H, Weaver M, Groopman J, Visscher B, Schwartz K, Detels R (1984) Quantitative changes in T helper or T suppressor/cytotoxic lymphocyte subsets that distinguish acquired immune deficiency syndrome from other immune subset disorders. Amer J Med 6:95–100
26. Lane HC, Masur H, Gelmann EP, Longo DL, Steis RG, Chused T, Whalen G, Edgar LC, Fauci AS (1985) Correlation between immunologic function and clinical subpopulations of patients with the acquired immune deficiency syndrome. Am J Med 78:417–422
27. Kalyanaraman VS, Sarngadharan MG, Robert-Guroff M, Miyoshi I, Golde D, Gallo RC (1982) A new subtype of human T-cell leukemia virus (HTLV-II) associated with a T-cell variant of hairy cell leukemia. Science 218:571–573
28. Gelmann EP, Popovic M, Blayney D, Masur H, Sidhu G, Stahl RE, Gallo RC (1983) Proviral DNA of a retrovirus, human T-cell leukemia virus, in two patients with AIDS. Science 220:862–865
29. Popovic M, Sarin PS, Robert GM, Kalyanaraman VS, Mann D, Minowada J, Gallo RC (1983) Isolation and transmission of human retrovirus (human T-cell leukemia virus). Science 219:856–859
30. Schupbach J, Popovic M, Gilden RV, Gonda MA, Sarngadharan MG, Gallo RC (1984) Serological analysis of a subgroup of human T-lymphotropic retroviruses (HTLV-III) associated with AIDS. Science 224:503–505
31. Sarngadharan MG, Popovic M, Bruch L, Schupbach J, Gallo RC (1984) Antibodies reactive with human T-lymphotropic retroviruses (HTLV-III) in the serum of patients with AIDS. Science 224:506–508
32. Barin F, M'Boup S, Denis F, Kanki P, Allan JS, Lee TH, Essex M (1985) Serological evidence for virus related to simian T-lymphotropic retrovirus III in residents of west Africa. Lancet 2:1387–1389
33. Kanki P, Barin F, Mboup S, Allan JS, Romet-Lemonne JL, Marlink R, McLane MF, Lee TH, Arbeille B, Denis F, Essex M (1986) New human T-Lymphotropic retrovirus related to simian T-lymphotropic virus type III (STLV-III$_{\text{AGM}}$). Science 232:238–243
34. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM, Hahn BH (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes (see comments). Nature 397: 436–441
35. Hirsch VM, Olmsted RA, Murphey Corb M, Purcell RH, Johnson PR (1989) An African primate lentivirus (SIVsm) closely related to HIV-2. Nature 339:389–392
36. Marlink R, Kanki P, Thior I, Travers K, Eisen G, Siby T, Traore I, Hsieh SS, Dia MC, Gueye EH, Hellinger J, Gueye NA, Sankalé JL, Ndoye I, Mboup S, Essex M (1994) Reduced rate of disease development with HIV-2 compared to HIV-1. Science 265:1587–1590
37. Kanki PJ, De Cock KM (1994) Epidemiology and natural history of HIV-2. AIDS 8:S1–S9
38. Kanki P, Peeters M, Gueye-NDiaye A (1997) Virology of HIV-1 and HIV-2. AIDS 2(Suppl B):S33–S42
39. Kanki P (1989) HIV-2 infection in West Africa. In: Volberding P, Jacobson M (eds) 1988 AIDS clinical reviews. Marcel Dekker, New York, pp 95–108
40. Kanki P, Meloni S (2009) Biology and variation in HIV-2 and HIV-1. In: Marlink RG, Teitelman S (eds) From the ground up: building comprehensive HIV/AIDS care programs in resource-limited settings. Elizabeth Glaser Pediatric AIDS Foundation, Washington, DC, pp 1–24
41. Katzourakis A, Tristem M, Pybus OG, Gifford RJ (2007) Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci USA 104:6261–6265
42. Wong-Staal F (1990) Human immunodeficiency viruses and their replication. In: Fields DM et al (eds) Virology. Raven, New York, pp 1529–1540
43. NIAID (2011) HIV Replication cycle. http://www.niaid.nih.gov/topics/HIVAIDS/Understanding/Biology/Pages/hivReplicationCycle.aspx
44. Jiang S (1997) HIV-1–co-receptors binding (letter; comment) (published erratum appears in Nat Med 1997 Aug; 3(8):817). Nat Med 3:367–368
45. Feng Y, Broder CC, Kennedy PE, Berger EA (1996) HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane G protein-coupled receptor. Science 272: 872–877
46. Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, Di Marzio P, Marmon S, Sutton RE, Hill CM, Davis CB, Peiper SC, Schall TJ, Littman DR, Landau NR (1996) Identification of

a major co-receptor for primary isolates of HIV-1. Nature 381:661–666

47. Cocchi F, DeVico AL, Garzino-Demo A, Cara A, Gallo RC, Lusso P (1996) The V3 domain of the HIV-1 gp120 envelope glycoprotein is critical for chemokine-mediated blockade of infection. Nat Med 2:1244–1247
48. Dittmar M, McNight A, Simmons G, Clapham P, Weiss R (1997) HIV-1 tropism and co-receptor use. Nature 385:495–496
49. Moore PS, Boshoff C, Weiss RA, Chang Y (1996) Molecular mimicry of human cytokine and cytokine response pathway genes by KSHV. Science 274:1739–1744
50. Cooper DA, Imrie AA, Penny R (1987) Antibody response to human immunodeficiency virus after primary infection. J Infect Dis 155:1113–1118
51. Loes S, de Saussure P, Saurat JH, Stalder H, Hirschel B, Perrin LH (1993) Symptomatic primary infection due to human immunodeficiency virus type 1: review of 31 Cases. Clin Infect Dis 17:59–65
52. Fauci AS, Pantaleo G, Stanley S, Weissman D (1996) Immunopathogenic mechanisms of HIV infection. Ann Intern Med 124:654–663
53. Pantaleo G, Graziosi C, Demarest JF, Butini L, Montroni M, Fox CH, Orenstein JM, Kotler DP, Fauci AS (1993) HIV infection is active and progressive in lymphoid tissue during the clinically latent stage of disease. Nature 362:355–358
54. Cavarelli M, Scarlatti G (2011) Human immunodeficiency virus type 1 mother-to-child transmission and prevention: successes and controversies. J Intern Med 270(6):561–579
55. Bongaarts J, Sinding S (2011) Population policy in transition in the developing world. Science 333:574–576
56. Bongaarts J, Pelletier F, Gerland P (2010) How many more AIDS deaths? Lancet 375:103–104
57. Root R (2010) Situating experiences of HIV-related stigma in Swaziland. Glob Public Health 5:523–538
58. Wu Z, Sullivan SG, Wang Y, Rotheram-Borus MJ, Detels R (2007) Evolution of China's response to HIV/AIDS. Lancet 369:679–690
59. Wang L (2007) Overview of the HIV/AIDS epidemic, scientific research and government responses in China. AIDS 21(Suppl 8):S3–7
60. UNAIDS (2010) Global report Europe and Central Asia. http://www.unaids.org/en/regionscountries/regions/easterneuropeandcentralasia/
61. Guillard EM, Eustache L (2007) Estimation de la séroprevalence du VIH en Haiti en 2007 selon le milieu de residence urbain et rural pour chacun des 10 départments. United States Agency for International Development, Washington, DC
62. Anonymous (2008) US Virgin Islands and Caribbean HIV epidemic need more attention, researchers say. HIV infection rate is high among sex workers. AIDS Alert 23:42–44
63. Bezemer D, de Wolf F, Boerlijst MC, van Sighem A, Hollingsworth TD, Prins M, Geskus RB, Gras L, Coutinho RA, Fraser C (2008) A resurgent HIV-1 epidemic among men who have sex with men in the era of potent antiretroviral therapy. AIDS 22:1071–1077
64. Mathers BM, Degenhardt L, Ali H, Wiessing L, Hickman M, Mattick RP, Myers B, Ambekar A, Strathdee SA (2010) HIV prevention, treatment, and care services for people who inject drugs: a systematic review of global, regional, and national coverage. Lancet 375:1014–1028
65. Strathdee SA, Hallett TB, Bobrova N, Rhodes T, Booth R, Abdool R, Hankins CA (2010) HIV and risk environment for injecting drug users: the past, present, and future. Lancet 376:268–284
66. Stockman JK, Strathdee SA (2010) HIV among people who use drugs: a global perspective of populations at risk. J Acquir Immune Def Syn 55(Suppl 1):S17–22
67. de Felipe B, Perez-Romero P, Abad-Fernandez M, Fernandez-Cuenca F, Martinez-Fernandez FJ, Trastoy M, Mata Rdel C, Lopez-Cortes LF, Leal M, Viciana P, Vallejo A (2011) Prevalence and resistance mutations of non-B HIV-1 subtypes among immigrants in Southern Spain along the decade 2000–2010. Virol J 8:416
68. Ryan CE, Gare J, Crowe SM, Wilson K, Reeder JC, Oelrichs RB (2007) The heterosexual HIV type 1 epidemic in Papua New Guinea is dominated by subtype C. AIDS Res Hum Retroviruses 23:941–944
69. Centers for Disease Control and Prevention (2011) High-impact HIV prevention – CDC's approach to reducing HIV infections in the United States. CDC, Atlanta
70. Katoff L, Dunne R (1988) Supporting people with AIDS: the Gay Men's health crisis model. J Palliat Care 4:88–95
71. Laga M, Galavotti C, Sundararaman S, Moodie R (2010) The importance of sex-worker interventions: the case of Avahan in India. Sex Transm Infect 86(Suppl 1):i6–7
72. Verma R, Shekhar A, Khobragade S, Adhikary R, George B, Ramesh BM, Ranebennur V, Mondal S, Patra RK, Srinivasan S, Vijayaraman A, Paul SR, Bohidar N (2010) Scale-up and coverage of Avahan: a large-scale HIV-prevention programme among female sex workers and men who have sex with men in four Indian states. Sex Transm Infect 86(Suppl 1):i76–i82
73. Tobian AA, Gray RH (2011) The medical benefits of male circumcision. J Am Med Assoc 306:1479–1480
74. Kreiss JK, Hopkins SG (1993) The association between circumcision status and human immunodeficiency virus infection among homosexual men. J Infect Dis 168:1404–1408
75. Weiss HA, Quigley MA, Hayes RJ (2000) Male circumcision and risk of HIV infection in sub-Saharan Africa: a systematic review and meta-analysis. AIDS 14:2361–2370
76. Siegfried N, Muller M, Deeks JJ, Volmink J (2009) Male circumcision for prevention of heterosexual acquisition of HIV in men. Cochrane Database Syst Rev: CD003362
77. Weiss HA, Hankins CA, Dickson K (2009) Male circumcision and risk of HIV infection in women: a systematic review and meta-analysis. Lancet Infect Dis 9:669–677
78. Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, Puren A (2005) Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial. PLoS Med 2:e298
79. Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F, Kiwanuka N, Moulton LH, Chaudhary MA, Chen MZ,

Sewankambo NK, Wabwire-Mangen F, Bacon MC, Williams CF, Opendi P, Reynolds SJ, Laeyendecker O, Quinn TC, Wawer MJ (2007) Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. Lancet 369:657–666

80. Wawer MJ, Makumbi F, Kigozi G, Serwadda D, Watya S, Nalugoda F, Buwembo D, Ssempijja V, Kiwanuka N, Moulton LH, Sewankambo NK, Reynolds SJ, Quinn TC, Opendi P, Iga B, Ridzon R, Laeyendecker O, Gray RH (2009) Circumcision in HIV-infected men and its effect on HIV transmission to female partners in Rakai, Uganda: a randomised controlled trial. Lancet 374:229–237
81. WHO/UNAIDS (2007) Announce recommendations about male circumcision as HIV prevention. Strategy should be employed with care. AIDS Alert 22:66–67
82. Shapiro RL, Hughes MD, Ogwu A, Kitch D, Lockman S, Moffat C, Makhema J, Moyo S, Thior I, McIntosh K, van Widenfelt E, Leidner J, Powis K, Asmelash A, Tumbare E, Zwerski S, Sharma U, Handelsman E, Mburu K, Jayeoba O, Moko E, Souda S, Lubega E, Akhtar M, Wester C, Tuomola R, Snowden W, Martinez-Tristani M, Mazhani L, Essex M (2010) Antiretroviral regimens in pregnancy and breast-feeding in Botswana. N Engl J Med 362:2282–2294
83. Chasela CS, Hudgens MG, Jamieson DJ, Kayira D, Hosseinipour MC, Kourtis AP, Martinson F, Tegha G, Knight RJ, Ahmed YI, Kamwendo DD, Hoffman IF, Ellington SR, Kacheche Z, Soko A, Wiener JB, Fiscus SA, Kazembe P, Mofolo IA, Chigwenembe M, Sichali DS, van der Horst CM (2010) Maternal or infant antiretroviral drugs to reduce HIV-1 transmission. N Engl J Med 362:2271–2281
84. World Health Organization (2010) Antiretroviral drugs for treating pregnant woman and preventing HIV infections in infants, Recommendations for a public health approach. 2010 revision.
85. Youle M, Wainberg MA (2003) Pre-exposure chemoprophylaxis (PREP) as an HIV prevention strategy. J Int Assoc Physicians AIDS Care (Chic) 2:102–105
86. Young TN, Arens FJ, Kennedy GE, Laurie JW, Rutherford G (2007) Antiretroviral post-exposure prophylaxis (PEP) for occupational HIV exposure. Cochrane Database Syst Rev:CD002835
87. Karim QA, Kharsany AB, Frohlich JA, Baxter C, Yende N, Mansoor LE, Mlisana KP, Maarschalk S, Arulappan N, Grobler A, Sibeko S, Omar Z, Gengiah TN, Mlotshwa M, Samsunder N, Karim SS (2011) Recruitment of high risk women for HIV prevention trials: baseline HIV prevalence and sexual behavior in the CAPRISA 004 tenofovir gel trial. Trials 12:67
88. Karim QA, Kharsany AB, Naidoo K, Yende N, Gengiah T, Omar Z, Arulappan N, Mlisana KP, Luthuli LR, Karim SS (2011) Co-enrollment in multiple HIV prevention trials – experiences from the CAPRISA 004 Tenofovir gel trial. Contemp Clin Trials 32:333–338
89. Abdool Karim Q, Abdool Karim SS, Frohlich JA, Grobler AC, Baxter C, Mansoor LE, Kharsany AB, Sibeko S, Mlisana KP, Omar Z, Gengiah TN, Maarschalk S, Arulappan N, Mlotshwa M, Morris L, Taylor D (2010) Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. Science 329:1168–1174
90. Grant RM, Lama JR, Anderson PL, McMahan V, Liu AY, Vargas L, Goicochea P, Casapia M, Guanira-Carranza JV, Ramirez-Cardich ME, Montoya-Herrera O, Fernandez T, Veloso VG, Buchbinder SP, Chariyalertsak S, Schechter M, Bekker LG, Mayer KH, Kallas EG, Amico KR, Mulligan K, Bushman LR, Hance RJ, Ganoza C, Defechereux P, Postle B, Wang F, McConnell JJ, Zheng JH, Lee J, Rooney JF, Jaffe HS, Martinez AI, Burns DN, Glidden DV (2010) Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. N Engl J Med 363:2587–2599
91. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, Hakim JG, Kumwenda J, Grinsztejn B, Pilotto JH, Godbole SV, Mehendale S, Chariyalertsak S, Santos BR, Mayer KH, Hoffman IF, Eshleman SH, Piwowar-Manning E, Wang L, Makhema J, Mills LA, de Bruyn G, Sanne I, Eron J, Gallant J, Havlir D, Swindells S, Ribaudo H, Elharrar V, Burns D, Taha TE, Nielsen-Saines K, Celentano D, Essex M, Fleming TR (2011) Prevention of HIV-1 infection with early antiretroviral therapy. N Engl J Med 365:493–505
92. Vissers DC, Voeten HA, Nagelkerke NJ, Habbema JD, de Vlas SJ (2008) The impact of pre-exposure prophylaxis (PrEP) on HIV epidemics in Africa and India: a simulation study. PLoS One 3:e2077
93. Veronese F, Anton P, Fletcher CV, DeGruttola V, McGowan I, Becker S, Zwerski S, Burns D (2011) Implications of HIV PrEP trials results. AIDS Res Hum Retroviruses 27:81–90
94. Alizon M, Wain-Hobson S, Montagnier L, Sonigo P (1986) Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. Cell 46: 63–74
95. Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, Deutsch P, Lifson JD, Bonhoeffer S, Nowak MA, Hahn BH, Saag MS, Shaw GM (1995) Viral dynamics in human immunodeficiency virus type 1 infection. Nature 373:117–122
96. Ho DD, Nuemann AU, Perelson AS, Chen W, Leonard JM, Markovitz M (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature 373:123–126
97. Starcich BR, Hahn BH, Shaw GM, McNeely PD, Modrow S, Wolf H, Parks ES, Parks WP, Josephs SF, Gallo RC, Wong-Staal F (1986) Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell 45:637–648
98. Wong-Staal F, Gallo RC (1985) Human T-lymphotropic retroviruses. Nature 317:395–403
99. Myers G, Pavlakis GN (1992) Evolutionary potential of complex retroviruses. In: Levy JA (ed) The retroviridae. Plenum, New York, pp 1–37
100. Goodenow M, Huet TH, Saurin W, Kowk S, Sninsky J, Wain-Hobson S (1989) HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions. J Acquir Immune Def Syn 2: 344–352
101. Wolfs TFW, deJong JJ, van den Berg H, Tijnagel JMGH, Krone WJA, Goudsmit J (1990) Evolution of sequences encoding the principal neutralization epitope of human immunodeficiency

virus type 1 is host dependent, rapid, and continuous. Proc Natl Acad Sci USA 87:9938–9942

102. Balfe P, Simmonds P, Ludlam CA, Bishop JO, Brown AJ (1990) Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. J Virol 64:6221–6233
103. Essex M, Kanki P (1997) Human immunodeficiency virus type 2 (HIV-2). In: Merigan T, Bartlett J, Bologenesi D (eds) Textbook of AIDS medicine, 2nd edn. Williams & Wilkins, Baltimore, pp 873–886
104. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH (2001) The origins of acquired immune deficiency syndrome viruses: where and when? Philos Trans R Soc Lond B Biol Sci 356:867–876
105. Gurtler LG, Hauser PH, Eberle J, von Brunn A, Knapp S, Zekeng L, Tsague JM, Kaptue L (1994) A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. J Virol 68:1581–1585
106. Simon F, Mauclére P, Roques P, Loussert-Ajaka I, Muller-Trutwin MC, Saragosti S, Georges-Courbot MC, Barre-Sinoussi F, Brun-Vezinet F (1998) Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat Med 4:1032–1037
107. Plantier JC, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemee V, Damond F, Robertson DL, Simon F (2009) A new human immunodeficiency virus derived from gorillas. Nat Med 15:871–872
108. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, Loul S, Ngole EM, Bienvenue Y, Delaporte E, Brookfield JF, Sharp PM, Shaw GM, Peeters M, Hahn BH (2006) Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science 313:523–526
109. Hemelaar J, Gouws E, Ghys PD, Osmanov S (2011) Global trends in molecular epidemiology of HIV-1 during 2000–2007. AIDS 25:679–689
110. Louwagie J, McCutchan F, Mascola J (1993) Genetic subtypes of HIV-1. AIDS Res Hum Retroviruses 9(Suppl 1): 147s–150s
111. Louwagie J, McCutchan F, Peeters M (1993) Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS 7:769–780
112. McCutchan F, Salimen MO, Carr JK, Burke DS (1996) HIV-1 genetic diversity. AIDS 10:S13–S20
113. Burke D, McCutchan F (1997) Global distribution of human immunodeficiency virus-1 clades. In: Vincent T, DeVita J, Hellman S, Rosenberg S (eds) AIDS: biology, diagnosis, treatment and preventions, 4th edn. Lippincott-Raven, Philadelphia, pp 119–126
114. Weniger BG, Takebe Y, Ou C-Y, Yamazaki S (1994) The molecular epidemiology of HIV in Asia. AIDS 8:13s–28s
115. Korber B, Brander C, Moore J, D'Souza P, Walker B, Koup R, Moore J, Haynes B, Myers G (eds) (1996) HIV molecular immunology database 1996. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos
116. Robertson D, Gao F, Hahn B, Sharp PM (1997) Intersubtype recombinant HIV-1 sequences. In: Korber B, Hahn B, Foley B, Mellors JW, Leitner T, Myers G, McCutchan F, Kuiken C (eds) Human retroviruses and AIDS 1997. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, pp III-25-III30
117. Jain MK, John TJ, Keusch GT (1994) Epidemiology of HIV and AIDS in India. AIDS 8:S61–75
118. Soto-Ramirez LE, Renjifo B, McLane MF, Marlink R, O'Hara C, Sutthent R, Wasi C, Vithayasai P, Vithayasai V, Apichartpiyakul C, Auewarakul P, Pena Cruz V, Chui DS, Osathanondh R, Mayer K, Lee TH, Essex M (1996) HIV-1 Langerhans' cell tropism associated with heterosexual transmission of HIV. Science 271:1291–1293
119. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM (2008) The challenge of HIV-1 subtype diversity. N Engl J Med 358:1590–1602
120. Vanharmelen J, Wood R, Lambrick M, Rybicki EP, Williamson AL, Williamson C (1997) An association between HIV-1 subtypes and mode of transmission in Capetown, South Africa. AIDS 11:81–87
121. Liitsola K, Holmstrom P, Laukkanen T, Brummer-Korvenkontio H, Leinikki P, Salminen MO (2000) Analysis of HIV-1 genetic subtypes in Finland reveals good correlation between molecular and epidemiological data. Scand J Infect Dis 32:475–480
122. Kalish ML, Korber BT, Pillai S, Robbins KE, Leo YS, Saekhou A, Verghese I, Gerrish P, Goh CL, Lupo D (2002) The sequential introduction of HIV-1 subtype B and CRF01 in Singapore by sexual transmission: accelerated V3 region evolution in a subpopulation of Asian CRF01 viruses. Virology 304:311–329
123. Ou CY, Takebe Y, Weniger BG, Luo CC, Kalish ML, Auwanit W, Yamazaki S, Gayle HD, Young NL, Schochetman G (1993) Independent introduction of two major HIV-1 genotypes into distinct high-risk populations in Thailand. Lancet 341:1171–1174
124. Herring BL, Ge YC, Wang B, Ratnamohan M, Zheng F, Cunningham AL, Saksena NK, Dwyer DE (2003) Segregation of human immunodeficiency virus type 1 subtypes by risk factor in Australia. J Clin Microbiol 41:4600–4604
125. Renjifo B, Fawzi W, Mwakagile D, Hunter D, Msamanga G, Spiegelman D, Garland M, Kagoma C, Kim A, Chaplin B, Hertzmark E, Essex M (2001) Differences in perinatal transmission among human immunodeficiency virus type 1 genotypes. J Hum Virol 4:16–25
126. Renjifo B, Gilbert P, Chaplin B, Vannberg F, Mwakagile D, Msamanga G, Hunter D, Fawzi W, Essex M (1999) Emerging recombinant human immunodeficiency viruses: uneven representation of the envelope V3 region. AIDS 13: 1613–1621

127. Hudgens MG, Longini IM Jr, Vanichseni S, Hu DJ, Kitayaporn D, Mock PA, Halloran ME, Satten GA, Choopanya K, Mastro TD (2002) Subtype-specific transmission probabilities for human immunodeficiency virus type 1 among injecting drug users in Bangkok, Thailand. Am J Epidemiol 155:159–168
128. Kanki PJ, Hamel DJ, Sankalé JL, Hsieh CC, Thior I, Barin F, Woodcock SA, Guèye-NDiaye A, Zhang E, Montano M, NDoye I, Essex ME, MBoup S (1999) Human immunodeficiency virus type 1 subtypes differ in disease progression. J Infect Dis 179:68–73
129. Kaleebu P, French N, Mahe C, Yirrell D, Watera C, Lyagoba F, Nakiyingi J, Rutebemberwa A, Morgan D, Weber J, Gilks C, Whitworth J (2002) Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. J Infect Dis 185:1244–1250
130. Kaleebu P, Ross A, Morgan D, Yirrell D, Oram J, Rutebemberwa A, Lyagoba F, Hamilton L, Biryahwaho B, Whitworth J (2001) Relationship between HIV-1 env subtypes A and D and disease progression in a rural Ugandan cohort. AIDS 15:293–299
131. Neilson JR, John GC, Carr JK, Lewis P, Kreiss JK, Jackson S, Nduati RW, Mbori-Ngacha D, Panteleeff DD, Bodrug S, Giachetti C, Bott MA, Richardson BA, Bwayo J, Ndinya-Achola J, Overbaugh J (1999) Subtypes of human immunodeficiency virus type 1and disease stage among women in Nairobi, Kenya. J Virol 73:4393–4403
132. Hu DJ, Vanichseni S, Mastro TD, Raktham S, Young NL, Mock PA, Subbarao S, Parekh BS, Srisuwanvilai L, Sutthent R, Wasi C, Heneine W, Choopanya K (2001) Viral load differences in early infection with two HIV-1 subtypes. AIDS 15:683–691
133. Sarr AD, Eisen G, Gueye-Ndiaye A, Mullins C, Traore I, Dia MC, Sankale JL, Faye D, Mboup S, Kanki P (2005) Viral dynamics of primary HIV-1 infection in Senegal, West Africa. J Infect Dis 191:1460–1467
134. Sagar M, Lavreys L, Baeten JM, Richardson BA, Mandaliya K, Chohan BH, Kreiss JK, Overbaugh J (2003) Infection with multiple human immunodeficiency virus type 1 variants is associated with faster disease progression. J Virol 77:12921–12926
135. Zhang M, Foley B, Schultz AK, Macke JP, Bulla I, Stanke M, Morgenstern B, Korber B, Leitner T (2010) The role of recombination in the emergence of a complex and dynamic HIV epidemic. Retrovirology 7:25–40
136. Ellenberger DL, Li B, Lupo LD, Owen SM, Nkengasong J, Kadio-Morokro MS, Smith J, Robinson H, Ackers M, Greenberg A, Folks T, Butera S (2002) Generation of a consensus sequence from prevalent and incident HIV-1 infections in West Africa to guide AIDS vaccine development. Virology 302:156–163
137. Essex M (2009) The impact of HIV variation on prevention and treatment. In: Kanki PM, Marlink RG (eds) A line drawn in the sand. Harvard University Press, Cambridge, MA, pp 231–242
138. Siegfried N, Uthman OA, Rutherford GW (2010) Optimal time for initiation of antiretroviral therapy in asymptomatic, HIV-infected, treatment-naive adults. Cochrane Database Syst Rev:CD008272
139. Spaulding A, Rutherford GW, Siegfried N (2010) Tenofovir or zidovudine in three-drug combination therapy with one nucleoside reverse transcriptase inhibitor and one non-nucleoside reverse transcriptase inhibitor for initial treatment of HIV infection in antiretroviral-naive individuals. Cochrane Database Syst Rev:CD008740
140. Spaulding A, Rutherford GW, Siegfried N (2010) Stavudine or zidovudine in three-drug combination therapy for initial treatment of HIV infection in antiretroviral-naive individuals. Cochrane Database Syst Rev:CD008651
141. Coffin JM (1996) HIV viral dynamics. AIDS 10:S75–84
142. Kanki P, Marlink RG (2009) A line drawn in the sand: responses to the AIDS treatment crisis in Africa. Harvard University Press, Cambridge, MA
143. United Nations (2001) Secretary-general urges united states business leaders to take concerted action against "Unparalleled Nightmare" of AIDS, U.N. Document
144. UNAIDS (2003) AIDS epidemic. data.unaids.org/publications/inc-pub06/jc943.epiupdate2003_en_pdf
145. PEPFAR (2011) United States President's emergency plan for AIDS relief. http://www.pepfar.gov/about/index.html
146. GFATM (2002) Global fund to fight AIDS, Tuberculosis and Malaria. http://www.theglobalfund.org/en/
147. BMGF (2011) Bill & Melinda Gates Foundation Global Health http://www.gatesfoundation.org/global-health/Pages/overview.aspx
148. CHAI (2011) Clinton Health Access Initiative. http://www.clintonfoundation.org/what-we-do/clinton-health-access-initiative
149. MAP (2011) World Bank Multi-Country HIV/AIDS Program (MAP). http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/AFRICAEXT/EXTAFRHEANUTPOP/EXTAFRREGTOPHIVAIDS/0,,contentMDK:20415735~menuPK:1001234~pagePK:34004173~piPK:34003707~theSitePK:717148,00.html
150. MSF (2011) Medecins sans Frontieres. http://www.msf.org/msf/about-msf/about-msf_home.cfm
151. Holmes C, Coggin W, Jamieson D, Mihm H, Savio P, Hope M, Ryan C, Moloney-Kitts M, Dybul M (2009) Measuring progress in reducing the costs of ARV drugs purchased by the president's emergency plan for AIDS relief, 2005–2007. In: Conference on retroviruses and opportunistic infections. Montreal, Canada
152. Shearer G, Clerici M (2010) Historical perspective on HIV-exposed seronegative individuals: has nature done the experiment for us? J Infect Dis 202(Suppl 3):S329–332
153. O'Brien SJ, Nelson GW (2004) Human genes that limit AIDS. Nat Genet 36:565–574

154. Wu X, Yang ZY, Li Y, Hogerkorp CM, Schief WR, Seaman MS, Zhou T, Schmidt SD, Wu L, Xu L, Longo NS, McKee K, O'Dell S, Louder MK, Wycuff DL, Feng Y, Nason M, Doria-Rose N, Connors M, Kwong PD, Roederer M, Wyatt RT, Nabel GJ, Mascola JR (2010) Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. Science 329:856–861
155. Koup RA, Graham BS, Douek DC (2011) The quest for a T cell-based immune correlate of protection against HIV: a story of trials and errors. Nat Rev Immunol 11:65–70
156. Bradac J, Dieffenbach CW (2009) HIV vaccine development: lessons from the past, informing the future. IDrugs 12:435–439
157. Girard MP, Osmanov S, Assossou OM, Kieny MP (2011) Human immunodeficiency virus (HIV) immunopathogenesis and vaccine development: a review. Vaccine 29:6191–6218
158. Fauci AS, Johnston MI, Dieffenbach CW, Burton DR, Hammer SM, Hoxie JA, Martin M, Overbaugh J, Watkins DI, Mahmoud A, Greene WC (2008) HIV vaccine research: the way forward. Science 321:530–532
159. Corey L, McElrath MJ (2010) HIV vaccines: mosaic approach to virus diversity. Nat Med 16:268–270
160. Caputo A, Gavioli R, Bellino S, Longo O, Tripiciano A, Francavilla V, Sgadari C, Paniccia G, Titti F, Cafaro A, Ferrantelli F, Monini P, Ensoli F, Ensoli B (2009) HIV-1 Tat-based vaccines: an overview and perspectives in the field of HIV/AIDS vaccine development. Int Rev Immunol 28:285–334
161. MacGregor RR, Boyer JD, Ugen KE, Lacy KE, Gluckman SJ, Bagarazzi ML, Chattergoon MA, Baine Y, Higgins TJ, Ciccarelli RB, Coney LR, Ginsberg RS, Weiner DB (1998) First human trial of a DNA-based vaccine for treatment of human immunodeficiency virus type 1 infection: safety and host response. J Infect Dis 178:92–100
162. Pantaleo G, Esteban M, Jacobs B, Tartaglia J (2010) Poxvirus vector-based HIV vaccines. Curr Opin HIV AIDS 5:391–396
163. Graham BS, Koup RA, Roederer M, Bailer RT, Enama ME, Moodie Z, Martin JE, McCluskey MM, Chakrabarti BK, Lamoreaux L, Andrews CA, Gomez PL, Mascola JR, Nabel GJ (2006) Phase 1 safety and immunogenicity evaluation of a multiclade HIV-1 DNA candidate vaccine. J Infect Dis 194:1650–1660
164. Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, Para MF (2005) Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. J Infect Dis 191:654–665
165. Gilbert PB, Peterson ML, Follmann D, Hudgens MG, Francis DP, Gurwith M, Heyward WL, Jobes DV, Popovic V, Self SG, Sinangil F, Burke D, Berman PW (2005) Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. J Infect Dis 191:666–677
166. Gilbert PB, Ackers ML, Berman PW, Francis DP, Popovic V, Hu DJ, Heyward WL, Sinangil F, Shepherd BE, Gurwith M (2005) HIV-1 virologic and immunologic progression and initiation of antiretroviral therapy among HIV-1-infected subjects in a trial of the efficacy of recombinant glycoprotein 120 vaccine. J Infect Dis 192:974–983
167. Buchbinder SP, Mehrotra DV, Duerr A, Fitzgerald DW, Mogg R, Li D, Gilbert PB, Lama JR, Marmor M, Del Rio C, McElrath MJ, Casimiro DR, Gottesdiener KM, Chodakewitz JA, Corey L, Robertson MN (2008) Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. Lancet 372:1881–1893
168. McElrath MJ, De Rosa SC, Moodie Z, Dubey S, Kierstead L, Janes H, Defawe OD, Carter DK, Hural J, Akondy R, Buchbinder SP, Robertson MN, Mehrotra DV, Self SG, Corey L, Shiver JW, Casimiro DR (2008) HIV-1 vaccine-induced immunity in the test-of-concept step study: a case-cohort analysis. Lancet 372:1894–1905
169. Gray G, Buchbinder S, Duerr A (2010) Overview of STEP and Phambili trial results: two phase IIb test-of-concept studies investigating the efficacy of MRK adenovirus type 5 gag/pol/nef subtype B HIV vaccine. Curr Opin HIV AIDS 5: 357–361
170. Nicholson O, Dicandilo F, Kublin J, Sun X, Quirk E, Miller M, Gray G, Pape J, Robertson MN, Mehrotra DV, Self S, Turner K, Sanchez J, Pitisuttithum P, Duerr A, Dubey S, Kierstead L, Casimiro D, Hammer SM (2010) Safety and immunogenicity of the MRKAd5 gag HIV type 1 vaccine in a worldwide phase 1 study of healthy adults. AIDS Res Hum Retroviruses 27(5):557–567
171. Voronin Y, Manrique A, Bernstein A (2010) The future of HIV vaccine research and the role of the Global HIV Vaccine Enterprise. Curr Opin HIV AIDS 5:414–420
172. Voronin Y, Phogat S (2010) HIV/AIDS: vaccines and alternate strategies for treatment and prevention. Ann NY Acad Sci 1205(Suppl 1):E1–9
173. Nitayaphan S, Pitisuttithum P, Karnasuta C, Eamsila C, de Souza M, Morgan P, Polonis V, Benenson M, VanCott T, Ratto-Kim S, Kim J, Thapinta D, Garner R, Bussaratid V, Singharaj P, El-Habib R, Gurunathan S, Heyward W, Birx D, McNeil J, Brown AE (2004) Safety and immunogenicity of an HIV subtype B and E prime-boost vaccine combination in HIV-negative Thai adults. J Infect Dis 190:702–706
174. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, Premsri N, Namwat C, de Souza M, Adams E, Benenson M, Gurunathan S, Tartaglia J, McNeil JG, Francis DP, Stablein D, Birx DL, Chunsuttiwat S, Khamboonruang C, Thongcharoen P, Robb ML, Michael NL, Kunasol P, Kim JH (2009) Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med 361:2209–2220
175. Kresge KJ (2009) Raft of results energizes researchers. IAVI Rep 13:4–5, 7–13, 17
176. Dieffenbach CW, Fauci AS (2011) Thirty years of HIV and AIDS: future challenges and opportunities. Ann Intern Med 154:766–771
177. World Health Organization (2010) Antiretroviral therapy for HIV infection in adults and adolescents: recommendations for a public health approach. 2010 Revision, Geneva

HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service

Douglass B. Lee, Jr., Kent M. Hymel
US Department of Transportation, Volpe National Transportation Systems Center, Cambridge, MA, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Characteristics of Managed Lanes
Evaluation of Pricing Alternatives
Future Developments
Note
Bibliography

Glossary

HOT Lane A limited access (expressway) separated highway service that permits qualifying HOVs to travel free, but allows non-HOVs to use the service for a fee (toll) [1]. Also HO/T.

Managed lanes Separated lanes that are operated or controlled actively in a way to limit congestion; charging a toll is one way to manage and an occupancy requirement is another. Other management tools include temporary shoulder use and dynamic signing (see http://ops.fhwa.dot.gov/freewaymgmt/managed_lanes.htm and http://managed-lanes.tamu.edu/).

Congestion pricing A toll or user charge on a facility that varies by time of day in accordance with demand, so as to maintain traffic flow on the facility at a freely flowing speed. Flat tolls that do not vary with demand, i.e., by time of day, do not qualify as pricing, and are not included within the scope of managed lanes.

HOV A high-occupancy vehicle, meaning there is at least one passenger in addition to the driver; the definition or criterion may require a minimum of two passengers (HOV 3+) in order to be eligible to use the HOV reserved lane.

Reserved lane A separated dedicated lane restricted to one or more special purposes for which not all vehicles are eligible. Bus lanes, HOV lanes, HOT lanes, truck lanes, and express lanes are forms of reserved lanes. Heavy vehicles may not be permitted at all on restricted lanes.

Toll Payment by a user vehicle to access a reserved lane; the toll may vary by time of day or with the level of demand so as to manage congestion and may also vary by vehicle type.

Expressway A design standard for multilane divided highways with grade-separated interchanges. HOT lanes are installed on mostly urban expressways.

Express lane A lane that allows some traffic to go faster than the default alternative, by restricting entry or skipping stops. A common usage refers to a tolled lane on an expressway, whether HOVs are discounted or not.

Express toll lanes (ETL) Restricted lanes that are tolled, perhaps with discounts for HOVs; facilities that charge discounted rates for HOVs are also referred to as hybrid systems.

General purpose lanes (GP) Lanes available for unrestricted traffic; also mixed flow lanes.

Slug lines Persons who line up at transit stops offering themselves as passengers to allow autos to meet the criterion for HOV eligibility in the same corridor.

Value pricing Application of the microeconomic theory of efficient pricing to the problem of highway congestion.

ORT Open road tolling, allowing vehicles with transponders to maintain highway speed and change lanes while the toll is collected electronically.

VOT The value of travel time, or the opportunity cost per hour of time spent in traveling, including congestion.

VOR The value of reliability, which is the consistency and predictability of travel times. Both speed and reliability are goals of managed lanes.

VMT Vehicle miles of travel.

Definition of the Subject

Economists have recognized for a long time (and transportation professionals more recently) that congestion on highways indicates that they are being operated inefficiently. The inefficiency stems from a failure to vary the price that drivers face with the level of demand for vehicle travel on the congested facility. In practice, congestion pricing on a transportation facility or service utilizes tolls that vary either by time of day, or directly with demand. Although there are many tolled highway facilities in the USA, including bridges, tunnels, and turnpikes, (see http://www.fhwa.dot.gov/ohim/tollpage.htm), few currently vary the toll by time of day (see http://ceprofs.tamu.edu/mburris/pricing.htm). One example is the New Jersey Turnpike and another is the Port Authority Hudson River crossings in New York City that have recently introduced a modest variation for peak periods [2].

Pricing one or more lanes of a facility so as to keep traffic on them moving smoothly, while leaving parallel lanes unpriced, is one way to increase the price toward a more efficient level. This strategy for congestion pricing is the dominant paradigm in the USA, and is referred to both as HOT lanes and as value pricing.

In most instances, tolled express lanes were preceded by lanes that were (or would have been) reserved for high-occupancy vehicles (HOVs). HOV lanes were created in an attempt to induce travelers to carpool by offering them a faster trip in return for the willingness to coordinate travel schedules and take only one vehicle. The goal can also be viewed as increasing passenger throughput by getting higher-occupancy vehicles to the head of the queue [3]. Although the utilization of HOV lanes is often well below capacity, reducing the criterion for high occupancy by one passenger may overshoot capacity by allowing too many vehicles. The concept of selling off the extra capacity to single-occupant vehicles (SOVs) was developed in order to "fine-tune" demand and more productively utilize the full capacity of the lanes. Such lanes became high-occupancy/toll lanes, or HOT lanes.

"Value Pricing" is a label that was developed for the purpose of avoiding the term "congestion pricing," thought at the time to create a negative impression. Subsequently, congestion pricing has been adopted as an acceptably neutral term, and value pricing has come to mean providing separated express service for an additional fee, on the same facility as the untolled service. Technically, such a differentiated alternative is a multiclass service, and it is this multiclass attribute that is the defining characteristic of HOT lane service.

Introduction

HOT lanes are the most popular form of road congestion pricing in the USA, but the form occurs rarely outside the USA. Other ways of controlling congestion, such as charging for entry to the downtown area of a city, are found in other countries but not in the USA. Thus HOT lanes should be seen within the context of a range of strategies for managing congestion by varying the price of road use so as to encourage traffic to avoid times and locations where demand is high. The HOT lane strategy is a blending of two strategies that involve offering motorists a choice: (1) the HOV strategy of offering speedier service to vehicles in which there is a minimum number of occupants, and (2) the pricing strategy of setting a toll on a facility that is high enough to maintain free-flowing traffic during peak periods, and lowering the toll at other times.

Each management strategy has its primary features and numerous variations in implementation and detail, and consequently each strategy has its own advantages and disadvantages. Although the general method of raising the price during the peak – congestion pricing – is clearly beneficial (if the cost of imposing the pricing is not too large), the costs and benefits of the various incomplete or partial methods of applying congestion pricing are not so clear. For that reason, the theory and practical methods for evaluating HOT lanes against other strategies for congestion management that use pricing are discussed. The issues are not fully resolved at either the theoretical or the practical levels, despite many strong assertions about the benefits of some of the strategies.

A source of the difficulty of rigorously evaluating HOT lanes is in their nature of offering two alternative "classes" of service for serving the same purpose, namely, a vehicle trip from point A to point B. Users are making several choices simultaneously, primarily, how many persons share the same vehicle, whether to travel during the peak period in the peak direction or avoid peak conditions, and whether to select the faster premium service or the lower-priced unrestricted service.

Because HOT lane management is entangled with the currently unsettled and highly contentious policy decisions of how to manage and pay for roads, the future evolution of HOT lanes and related congestion pricing schemes is uncertain at this time in the USA.

Characteristics of Managed Lanes

The discussion that follows covers three main topics: what are HOT lanes and other "managed" lanes, how can we tell whether they make us better off, and how will they evolve? The first topic – description – outlines the characteristics of managed lanes within the urban transportation context. The second topic – evaluation – explains the economic theory for assessing the benefits of the HOT lane innovation and its alternatives. The third topic – future developments – speculates about the likely outcomes of conflicting current trends.

Distinguishing Features of Priced Lanes

HOV lanes, HOT lanes, and express lanes have been generically called "managed" lanes, implying that certain attributes of the service are controlled by the provider so as to affect the quality of service. The purpose of "lane management" is to reduce or control congestion on the lanes that are managed. Pricing is one mechanism for managing highway lanes, but occupancy requirements and truck-only lanes are also considered to be tools for lane management (Collier and Goodin [4] survey the planning, design, and operation of managed lanes, offering several case studies).

Although pricing is often classified as demand management, express lane management is more accurately viewed as supply management, even to the point of constructing additional lanes. One planning study listed HOV lanes, HOT lanes, express toll lanes, and truck-only toll lanes, and "special use lanes such as express, bus-only, or truck-only" as the full range of "management tools" [5]. To date, the multiclass service strategy for dealing with congestion has been applied almost exclusively in the USA (see, e.g., [6] on European experience).

HOT lanes have been utilized in metropolitan areas on limited access urban expressways. Table 1 lists projects currently in operation; others are under construction or planned. An express lane can be a single lane per direction or several; two lanes per direction is common and also the maximum in practice.

Electronic Toll Collection Electronic toll collection (ETC) on HOT lanes is required in practice, and can be implemented using prepaid accounts, automatic debit accounts, or via video license plate readers [11]. The toll paid to use the express lanes may or may not apply to HOVs, but the price is set on the tolled vehicles so as to maintain free-flowing conditions on the express lanes (the SOV share gets as high as 16% in Minneapolis [12]). Congestion on the remaining general purpose (GP) lanes is left to fall where it may. Users are usually informed of the reserved lane toll near points of

HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Table 1 Existing HOT/Express lanes in the USA in 2010

Location	Name	Year opened	Length (mile)	Lanes (reversible)
Orange Country, CA	SR-91 Express lanes	1995	10	4
Houston, TX	QuickRide (US 290)	2000	13	1R
Houston, TX	Katy freeway managed lanes	2009	13	1R
Denver, CO	I-25 HOT lanes	2006	7	4
San Diego, CA	I-15 Fastrak	1996	8	4R
Miami, FL	I-95 HOT lanes northbound	2008	21	2
Minneapolis, MN	I-394 MnPASS express lanes	2005	13	2
Minneapolis, MN	I-35 MnPASS express lanes	2010	13	2
Salt Lake City, UT	I-15 HOT lanes	2006	38	2
Seattle, WA	SR-167 HOT lanes pilot	2008	9	2

Sources: [7–10].
The number of lanes and their reversibility often changes from one section of the facility to another

entry, especially when tolls are determined by concurrent congestion levels. Implementation of a HOT lane facility in practice is a demanding task. There are many reports that describe the process of planning, outreach, consensus, physical configuration, barrier separation, access points, operational details, etc. Two extensive guidebooks are Parsons Brinkerhoff [13] and PB Study Team [14].

Facilitating payment by manual toll collection, exact change, credit card, or other procedure that requires the vehicle to come to a full stop is very costly and disruptive. The trend is toward open-road tolling (ORT) in which the toll is collected electronically without any slowing down, and all HOT lanes currently operate as ORT. Mixed toll collection requires that non-ETC payers be segregated from the main travel stream, with enough storage space provided to serve them off-line, and then merged back into the free-speed stream.

Users without transponders can be billed by reading their license plates via video camera and sending a bill to the owner of the plate. The 407 highway around Toronto, Canada, has been primarily ETC for over a decade, with supplemental video tolling for the small share of traffic without transponders [15].

The HOV Lane Precursor Tolled express lanes likely would not have occurred in the USA without the preexistence of HOV lanes (although the first – SR-91 – was built directly rather than being a conversion). Ironically, perhaps, separate lanes for high-occupancy vehicles was devised as an incentive mechanism that would achieve the behavior objectives of pricing without its deterrent effect.

The concept of giving priority to vehicles having more than one passenger was implemented primarily as a means to encourage bus transit and carpoolers over SOV (single occupant vehicles), and thus reduce the number of vehicles needed to carry the same number of passengers. The physical configurations to create the separate service range from lanes marked with a diamond but no barrier, to concrete barrier separation, to reversible lanes with barriers, queue bypass lanes, and separated bus-only lanes [16]. Lanes may be restricted only during weekday peaks or during all hours and days.

The potential gain from providing HOV lane service is a reduction in the number of vehicles as a result of ridesharing. The magnitude of the benefit is greater if existing carpools are not reduced, if new carpools are formed from existing peak travelers, if bus travelers do not switch to auto, and if the increase in vehicle occupancy is significant. In practice, the sources of additional occupants are same family (75%), work colleagues (20%), and casual carpooling [17]. The last depends heavily on the physical configuration of the corridor and the level of transit service (“slugs” come from transit). Thus the level of actual induced carpooling tends to be low, and not enough to overcome the handicaps of lower capacity overall and particularly on the HOV lane (Poole [18] notes the low effectiveness of occupancy requirements; Poole and Orski [19] describe the historical backlash against HOV lanes and the origins of HOT lanes; Lee [20] describes the loss of capacity). Because not all HOV operators provide data on utilization rates or violation rates, actual HOV performance can be difficult to assess [21].

The number of HOV facilities (roughly 350 in the USA) greatly exceeds the number of HOT lane examples. Some HOV lanes are over 50 miles in length (see [22] for a full HOV inventory and details). The large number of existing HOV lanes indicates that the pool of potential conversions from HOV to HOT lanes is large. Because the user cost of assembling a carpool is a fixed cost, per trip, the incentive is strongest for long commuting trips (over 35 miles, according to [23]). Conditions suitable for conversion of HOV lanes to HOT lanes are analyzed in detail in [24].

Multiclass Service A possible outcome of applying either the microeconomic theory of efficient pricing or of congestion management could be a strategic decision to offer two (or more) types of service whose primary differentiating characteristic is speed. The criteria for eligibility for the higher quality service could be a minimum vehicle occupancy level and/or the payment of a toll that varies with the level of demand.

The success and benefits of such a multiclass strategy depend upon heterogeneity in the values of travel time among the group of users at the time the service is offered. If all vehicles have the same value of travel time (VOT), then multiclass service may not be any better than single class service and could be worse if there are costs in providing the service. Hence the desirability of offering two-class service depends upon the distribution of the values of (vehicle) time in the traffic stream. Ideally, there would be two relatively distinct groups of vehicles, with respect to time value. The emphasis on valuation at the vehicle rather than the person level is that it is the vehicle that is consuming the scarce resource, and different occupancy levels could provide sufficient distinction among vehicles to warrant multiclass service even if all travelers have the same VOT (see Both Services Priced Efficiently on p. 27).

Eligibility Requirements Use of the special lane is controlled by vehicle occupancy requirements, price, and other user or vehicle characteristics. Usually only one requirement (occupancy or price) needs to be satisfied. Other characteristics (proposals have been made to include low-emission vehicles (LEV) and hybrids as also deserving favorable treatment) are unrelated to congestion management but have taken a significant share of capacity on some facilities.

The effectiveness of HOV (untolled) lanes is difficult to sustain because fluctuations or growth in demand necessitate changes in the required number of vehicle occupants in order to insure free-flow conditions on the reserved lanes. If congestion occurs frequently with a two-person criterion for occupancy, then the criterion can be raised to three persons. But such a discrete change may result in the lanes being underutilized, at least for some period of time. In general, HOV lanes have not been shown to be effective in increasing occupancy or reducing congestion [25], compared to the same number of lanes in GP service. So far, HOV and HOT lanes have been the result of new construction, which inevitably reduces congestion below what it would have been without the added capacity.

Pricing refers to the application of microeconomic concepts for reconciling supply and demand, when normal market responses (changing price and quantity) are nonexistent or too rigid (other than queuing). For a given capacity (i.e., in the short run), price can be set so as to clear the market without resorting to queuing. Microeconomic theory requires setting price equal to marginal cost in order to achieve efficient utilization of the scarce resource (capacity), but “management” can be applied more broadly to achieve other appropriate goals such as a level of service (LOS) standard.

Performance Objectives Because the incentive for inducing vehicles to rideshare or pay the price depends upon a promise of faster and more reliable service, management of multiclass services places a primary emphasis on ensuring that the exclusive service is always superior. HOT lanes are sustainable in the short run and medium run because the price on the express lanes can be adjusted in time increments as short as 3 min, so as to maintain a consistent level of service (LOS) (the 3 min adjustment rate is claimed for the MnPASS express lanes in Minneapolis [12]).

The standard for level of service (LOS) can be defined in several ways, including the ratio of volume to capacity or the average speed for a given type of facility. Liu et al. [9] citing Supernak et al. [26] determine that US HOT lanes are priced to achieve 45 mph 90% of the peak period (approximately LOS C), which is congested but still flowing smoothly (see Transportation Research Board [27]). In practice, HOT lane prices can be adjusted upward and sometimes downward to maintain service standards; for example, tolls on the SR-91 express lanes in southern California were actually lowered in 2009 following a recession-caused decrease in demand [28]. Whether the tolls will be sufficient to maintain and replace the facility over the long run is largely unknown at present because the history of such lanes is still fairly short (see Revenues on p. 15 below).

Additional Costs of Multiclass Service

Creating a two-class (or more) service incurs several costs that can be avoided or lessened if the facility is left in a single class of service.

Safety and Enforcement With separate but unequal services, there is an incentive to take advantage of the faster service if that can be done without paying the price, whether the price is the inconvenience of carpooling or in the form of a toll. The incentive is especially strong if the GP lanes are barely moving while the restricted lanes are underutilized. If there are no barriers between the lanes, then some drivers may be tempted to switch lanes. Because this places fast and slow vehicles in juxtaposition, changing lanes creates a safety problem. If the violation rates are high (15%) for a tolled service, there is an additional incentive to switch lanes, and there is a revenue loss to the owner. A study found a 15% violation rate on a Phoenix HOV lane that was (as is common) not barrier separated, even with a fine of $350 [29]. Finally, if eligibility includes a minimum number of occupants in the vehicle (or a hybrid gasoline-electric vehicle), enforcement may be more costly because visual verification of the occupancy or vehicle make is required (Poole [18] reviews the technology for occupancy verification and mentions that where HOV2 are eligible for free travel they are a majority of vehicles).

To minimize the cost of enforcement, physical separations of some kind are needed. There is thus a tradeoff between the cost of enforcement and the cost of physical infrastructure. Whatever the optimum balance, the net result is an increase in cost over a single undifferentiated service.

Accidents tend to be higher on HOV/HOT facilities, affected by differing traffic speeds when merging narrower lane width and shoulder width, but not affected by continuous versus limited access [30].

Infrastructure HOV services require shoulders, transition, or merging lanes, breakdown and emergency services, and other features that may be needed above those required for GP service. Some HOV lanes are simply marked with diamonds, with no buffer or barriers or pylons demarcating the lanes. Vehicles can enter or leave at any point.

Transforming HOV into HOT additionally requires toll collection gantries and machines, and back-office functions related to user accounts. Toll administration may or may not impose greater costs than normal ETC.

Electronic signs and other information systems are used to inform users of the toll rate on the express lanes, at the time the user needs to make a choice. Other information may also be conveyed, such as the days or hours of operation.

Loss of Capacity In general, the more lanes moving in the same direction, the greater the capacity per lane. This phenomenon occurs because vehicles can weave across lanes to make use of space in front of slower or less aggressive drivers. Hence, the disruption from entrance and exit ramps is lessened. But with barrier-separated lanes, such opportunities for weaving are diminished, which results in an effective loss of capacity. Rough averages for capacity and speed for different number of contiguous lanes are shown in Table 2. Actual capacities and speeds vary substantially from these upper-bound estimates on capacity and speed ranges. As for usage, many single HOV lanes rarely get above 1,000 vehicles per hour, perhaps as a result of the uneven demand from eligible vehicles.

One or two lanes is the norm for express lanes, whereas GP lanes or full facility pricing can apply across all numbers of lanes. Capacities are often dependent on the sources for inflow (feeder links) and recipients for outflow (exits).

Cassidy et al. [31] argue that carpool lanes can smooth traffic flows at bottlenecks, in all lanes including GP lanes, but these conditions apply when volume is sufficient to a cause breakdown in flow.

Imbalanced Throughput The maximum flow on the facility occurs when all lanes are part of the same flow. When some lanes are allowed to go faster, the total throughput for the two services combined is less than for the same number of lanes without separation. This result is due to the nonlinearity of delay: The marginal delay from adding a vehicle to a congested stream is always greater than the marginal reduction in delay from taking the vehicle from a less congestion stream. While popular intuition is perhaps as often wrong as right, the effect of unbalanced flows per lane is perceived as making throughput worse [32].

HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Table 2 Estimated capacity and speed by number of lanes in the same direction [27]

Number of lanes	Capacity (vehicles per lane per hour)	Free-flow speed (mph)
1	1,600 [17]	45–55
2	1,800	58 Goodall and Smith [12] claim average speeds as high as 78 mph on a two-lane HOT lane service in Minneapolis
3	2,050	65
4 or more	2,200	70

If the objective is to minimize overall vehicle delay, the optimal result occurs when traffic is distributed equally across all lanes; at that point, the reassignment of any vehicle to another lane increases delay more on that lane than it reduces delay on the lane it left. This is necessarily the case because the second derivative of the delay curve is always positive. As an example, the speed–volume relationship can be represented by a curve of the BPR form

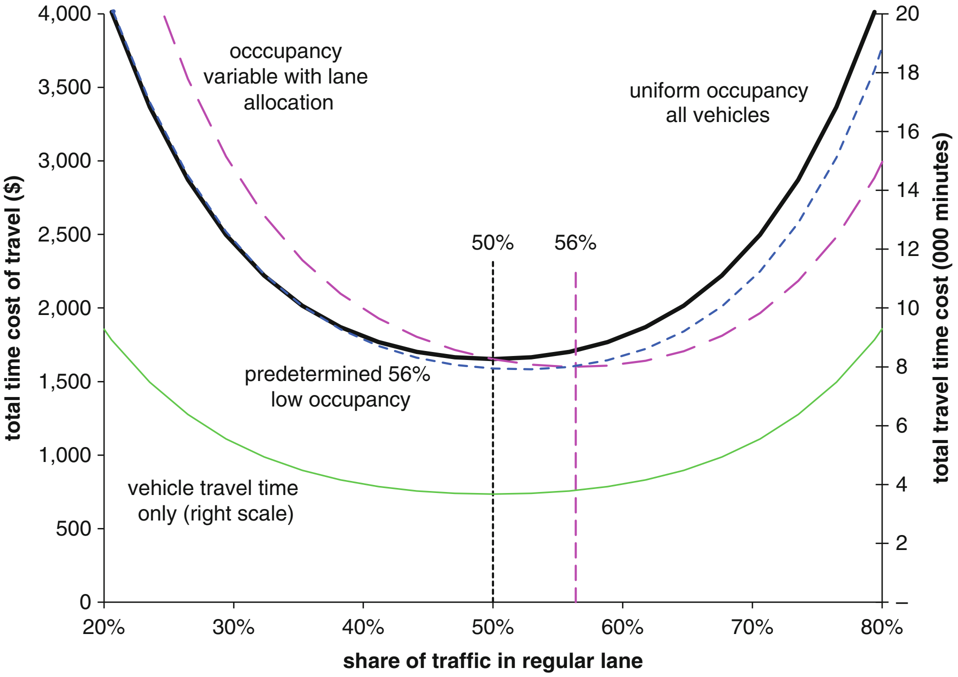
$$S(v) = \frac{F_s}{1 + \alpha\left(\frac{v}{c}\right)^{\gamma}}$$

where $S(v)$ = speed as a function of volume, F_s = free-flow speed, c = capacity, and α and γ are parameters [33]. Travel time is the reciprocal of speed.

This function is used in the illustrative example shown in Fig. 1, with $\alpha = 0.3$ and $\gamma = 7.0$. Consider a two-lane facility with a fixed vehicle flow (one direction), where the problem is how to allocate traffic to each lane to minimize total travel time. Proceeding from left to right in the diagram, as more traffic is added to lane 1 the total time cost declines because the extreme congestion in lane 2 is relieved without heavily impacting lane 1. The lowest curve shows vehicle delay for all vehicles combined. The minimum is reached when half the traffic is assigned to each lane.

Converting time into cost, through vehicle occupancy and the individual value of travel time (VOT), produces the thick solid curve that replicates the time curve (but higher on the graph) so long as both occupancy and VOT are the same for all vehicles.

If the number of occupants and their VOT are not uniform, lower travel cost may be obtained by sorting users into categories of higher occupancy and higher time value, rewarding the latter with faster travel times. Taking the numbers shown in Table 3, and assuming that the occupancy and time values are determined by the lane to which the vehicle is assigned, the time cost by the share assigned to lane 1 is shown by the heavier dashed magenta curve. VOT per person is the same for all travelers, but occupancy is twice as high on lane 2 as on lane 1. The optimum (least total time cost) occurs when about 56% of the traffic is on lane one, and the total cost is slightly lower than the average in part because higher VOT vehicles go faster and in part because there are fewer passengers (56% of the vehicles



HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Figure 1
Total travel time cost as a function of traffic share allocated to lane 1

have 1.2 passengers versus 50%). This curve crosses the uniform-traveler (dotted) curve at 50% because at that point they are the same.

If the number of vehicles in each category is determined a priori, such as in the right-hand column of Table 4, in this case at approximately 50%, the cost is shown by the heavy blue curve, which has a minimum between 50% and 56%. At the latter allocation, it is the same as the previous curve in which occupancy depended upon the lane to which the vehicle was assigned. Again, the parameters are the same at that point. Assigning more than 56% to lane 1, however, means assigning high-occupancy vehicles to the slow lane (more people at slower speed), so total cost rises rapidly. To the left of 56%, there are fewer people in the constant-occupancy scenario than the variable occupancy scenario, so costs are a little lower.

Changing VOT has the same effect as occupancy, which is to make the time of some vehicles higher valued than others. For a difference of a factor of 2.0 in the vehicle value of time, the least-cost flow still only varies from the equally balanced flow by less than five percentage points.

HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Table 3 Assumed average vehicle occupancy (AVO) for regular GP and HOT lanes

Vehicles	Occupancy	VOT	No. veh.
Low occupancy/regular	1.2	$15.00	1,900
High occupancy/HOT	2.4	$15.00	1,500
Average (unweighted)/ total	1.8	$15.00	3,400

HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Table 4 Seattle SR-167 single lane speed and volume (Lin et al. [9])

Segment No.	% Peak flow <800 vph	% Flow >1,050 vpl %Speed >55 mph	Flow <1,050 vpl
1	82	1	99
2	31	7.2	92
3	9	10	77.4
4	21	11.1	87
5	8	11.2	87.7
6	30	7.5	91.4

The Los Angeles County MTA has announced the intention of converting HOV lanes to HOT without reducing service on the GP lanes, which should not be a problem; the problem occurs if a GP lane is converted directly to HOT. So long as the HOV lane is a new lane, the political backlash is mild compared to taking an existing lane out of service. When the same capacity is split into classes, flow in the GP lanes may exceed the breakdown level, resulting in "hypercongestion," a condition in which capacity is less than what the facility is capable of under ideal conditions. The extent and magnitude of this phenomenon is open to question [34], but at a network level an overloaded system with unbalanced flows appears to be less productive than one that is not unbalanced [35].

Utilization Despite the pricing, HOT lanes may still be underutilized, if there is general resistance to paying tolls. Perhaps unusual, the HOT lanes in the Seattle area appear to be underutilized, despite congestion on the GP lanes. The data in Table 4, inferred from volume and occupancy detectors, indicate difficulty in achieving full utilization of all lanes.

Time savings on SR-167 averaged 9.1 min over 11 miles, but can get as high as 20 min. The express lanes consistently operate over 45 mph, with a mean travel time of 8.5 min and a standard deviation of 26 sec. The HOT service is relatively new, and demand is continuing to grow.

Currently, the existing ten US HOT lane examples are able to effectively manage the level of service on the HOT lanes by changing prices dynamically or by modifying fixed prices based on observation of traffic flows and patterns. Diagnostic guidance on obtaining close to optimum utilization in HOV or HOT lanes is provided in [36].

Demand

It is well known that express lane users do not always choose to pay for the faster service whenever they use the facility [37]. A large share of users pay the toll two or three times a week and use the GP lanes on other days, making the choice to use the express lanes trip by trip. Hence the value of saving travel time varies considerably for the same individuals from one day to another, as well as among travelers.

Although peak travelers moving in the peak direction on the same facility are relatively affluent compared to the general population, they are far from identical. Occupancy, of course, also makes a difference in the vehicle's willingness to pay for faster speed.

Some demand studies exist for predicting the factors that determine which class of service a driver will select on the basis of traveler attributes, but less work exists that examines how variations in service attributes affect lane choice. Most new HOT lane users were previously SOVs in the GP lanes on the same facility (or HOV2s), at least in the case of conversions from HOV to HOT service [38]. Relative to other peak travelers, express lane users have significantly higher income and education than the average, even of other peak highway users. The small share of lower-income toll payers is likely to be in carpools. For the most part, carpoolers use the express lanes because they are already carpooling; few carpools are formed in order to use the lanes. While there are no surprises here, it may be reassuring that HOT lanes do not cannibalize carpools or transit.

Market Research on Express Lane Users The demand model may be approximated as this: Users estimate the magnitude of the time savings and decide whether the price is right, without changing mode or occupancy. Other reasons as well, however, factor in. Drivers in Minneapolis use the HOT lanes even when there is no time advantage, perhaps by (previous) habit or perhaps to ensure reliability [12].

Zmud et al. [39] find toll willingness to pay is affected by income, age, trip purpose, time of day, trip distance, and time savings (time of day is, of course, a surrogate for demand level).

Burris et al. [40] find that tolling HOVs causes a small but significant reduction in their numbers, while increasing total revenue.

Revealed Preference on Toll Lane Users Aggregate demand tends to grow, due to rising income, population, and economic activity. Liu et al. [9] observe that a higher-than-typical price at a given time may signal to users that the express lanes may be approaching congestion due to high overall demand, hence an indication that the GP lanes are slow. For SR-167 in Seattle, a toll of $1–2.00 for a speed increase from 30 to 60 mph over 11 miles yields a time saving of around 11 min, for an implied vehicle value of time (VOT) of less than $10 per hour. The toll rate is adjusted every 5 min. A good deal of data has been collected on user behavior regarding SR-91, including interview surveys, and these data have been extensively studied [41].

Diversion For HOT lanes, most of the diversion is between the two classes of service. Because the lanes have resulted in additional capacity when the multiclass service was initiated, there has not been much need for concern about diversion to other facilities.

If, however, existing lanes are converted to express or HOT lanes, the overall price on the facility will rise (in dollar terms for the tolled lanes, and in delay on the GP lanes), so diversion to other facilities will need evaluation. Depending upon the substitutability of nearby arterial streets – measured in travel time – tolls may need to be lowered on the express lanes in order to retain enough traffic to prevent arterials from being choked.

This problem is different in kind from the consideration of delay on the GP lanes when tolls are raised on the tolled lanes. If the facility manager is responsible for ameliorating delay on the GP lanes, the level of service on the express lanes may need to be compromised to retain reasonable flow on the GP lanes [42, 43]. This is how the I-95 Express Lanes in Miami are managed, with the project manager coining the term "quality throughput." Pricing both services, of course, would allow for much greater control and less compromise.

Variations in the Value of Travel Time Savings Variation in time values is widely asserted, largely on circumstantial grounds (e.g., toll payers vary from day to day in their choice). Patil et al. [44] use stated preference (SP) to find a 95% confidence range of 41–79% of the wage rate, a fairly broad band. Small, Winston, and Yan [45] use both stated preference and revealed preference data to determine that motorists place high value on travel time and travel time reliability, with considerable heterogeneity in both.

Travel Time Reliability Express lanes offer significantly more reliable travel times, as well as shorter travel times. The general pattern from user surveys and potential user surveys is that most travelers are occasional users who decided each day how urgent their arrival time is [46]. Using data from SR-91 in southern California, Brownstone and Small [47] find high values for both time and reliability. Defining and measuring reliability is a challenging task, and the real or perceived benefits have not been integrated into benefit-cost evaluations [48, 49]. Patil et al. [50] surveyed travelers on the Katy Freeway in Houston regarding their travel on the GPLs and the managed lanes. One area of investigation was the travelers' willingness to pay for a trip on the managed lanes under unusual situations – basically when the traveler was in more of a hurry or pressed for time, for example, running late for a meeting or when they see bad weather that may cause congestion. They found the willingness to pay for travel time savings in those instances was between two and four times as much as regular travel time savings.

Evidence from travelers using the express lanes on I-394 in Minnesota also points to express lane users placing a high premium on travel time reliability. Cho et al. [51] obtained data on travel times for the GPLs and ELs plus toll collection data from every transaction in 2008. Based on these data it was found that approximately 50% of the paying EL users paid for travel time savings of 1 min or less. Mostly due to these small time savings, the average values of time were found to be considerably higher (ranging from $73 to $116 per hour) than what was found in the literature (often around $10–25 per hour). This would indicate a willingness to pay for something other than the travel time savings offered by the lane.

Revenues

A policy that both reduces congestion and raises revenues should be an attractive policy, but that has not made congestion pricing in general widely adopted. While some commentators express the observation that the revenue potential seems to have overshadowed the congestion management benefits, the evolution of the HOT lanes concept from HOV to allowing a few more vehicles to "buy in" to the service does not suggest that revenue raising is the primary motivation.

Nonetheless, tolls do generate revenue, and more under some circumstances than others. Tolls used for congestion management vary with demand and hence by time of day. Flat tolls that do not vary with time of day are primarily if not exclusively for revenue generation to pay for the road. Such tolls may vary by vehicle type to reflect costs of serving heavier or larger vehicles, but the rates are generally not set so as to obtain a particular mix of vehicles.

Monopoly The monopolist in microeconomic theory faces a downward sloping demand curve (unlike a firm in a competitive market) and sets prices that are intended to maximize profit. The monopolist does not necessarily have full control over the market price and does not face totally inelastic demand, but any degree of downward sloping demand for the firm's output conveys some measure of monopoly power. In theory, the monopolist maximizes profit by setting price on the demand curve at the quantity where marginal cost equals marginal revenue.

If profit is positive at any volume (it might not be, depending upon the cost structure), the MC = MR point will yield the highest profit. More generally, the monopolist may not choose to seek the absolute maximum profit in the short run, perhaps lowering the price to increase market share or provide a cushion of loyal customers. Express lanes are typically priced so as to reliably maintain 45 mph speeds during the peak periods, which is normally a higher volume than the profit maximizing volume.

The degree of monopoly power possessed by toll roads or express lanes is controversial. There is inevitably some spatial monopoly, but given the ability of travelers to carpool, switch to transit, take an alternative route or destination, shift to an off-peak time, or forgo the trip, the demand for any single facility is necessarily somewhat elastic.

Exemptions A much more important factor with respect to revenue is the level of exemptions from paying the toll. If two-person carpools go free, it is possible to use all the capacity for HOVs and leave no paying customers. Adding exemptions for electric vehicles and hybrids further erodes the revenue yield. Broadly speaking, as the share of nonpaying vehicles increases, the revenue production of the facility becomes a less viable purpose for the express service [52].

The significance of revenue, then, is whether it can pay for the operation of the facility and the collection of the revenue. Varying the toll according to demand and billing each vehicle separately is more costly than collecting the same revenue through a gas tax, so the question of whether the service is a good idea depends upon the benefits of better congestion management in travel time savings, reliability, and providing the option to avoid the GP lanes, compared to the costs of creating and maintaining a more accurate pricing system.

HOV discounts or exemptions are problematic because compliance is hard to verify. This is true for enforcement of all HOV lanes, but the addition of tolling certain vehicles while others go free greatly complicates the task.

Price Volatility

Express lanes are those which charge all users a toll, perhaps discounted for HOVs, while HOT lanes allow vehicles meeting the HOV criterion to travel free. SR-91 charges HOVs half price. A problem with exempt or discounted vehicles is that it creates uncertainty about the potential revenue stream and amplifies any volatility in the price per trip over the time of day or day of the week. As the share of traffic that is insensitive to price (free or discounted) becomes larger, the remaining price-sensitive traffic has to respond proportionately more to maintain the congestion at LOS C. At the limit, when the share of price-sensitive traffic approaches zero, the fluctuations in price needed to maintain LOS C become infinite. Poole [53] describes an Atlanta planning study in which the HOT facility would generate no revenues in the peak period at some time in the future, because the capacity would be fully used by HOV2+ vehicles and paying vehicles would be entirely excluded. Under such circumstances, the investment of private capital is unlikely without explicit guarantees. The requirements for exemption can be increased (e.g., from 2+ to 3+), causing a sudden drop in the toll for SOVs (and now two-person carpools), but this volatility does not provide confidence in the long-term stability of the operation.

Another possible source of volatility is the misperception by some express lane users that the price is a measure of congestion on the GP lanes, rather than a price based on the number of express lane users. Miami-area HOT lane users, among others, have expressed the belief that the price is an index of congestion on the free lanes [54]. While the perception has a perverse effect on the use of price as a rationing instrument, it indicates that express lane customers are willing to pay a premium for guaranteed higher quality, or more specifically, reliability. Posting the time savings between the two services could serve as a counter to the misperception.

Cost Recovery Because no revenues are collected other than fuel excise taxes, HOV lanes are not held to any financial recovery standard.

If express lanes are to be regarded as a means for financing transportation infrastructure, the tolls will need to be higher than currently, apply to all vehicles, and apply to a large share of daily traffic [55]. There may be some circumstances where the surplus revenue after paying for operation and toll collection is large enough to recover the capital cost of the lane and associated infrastructure (barriers, gantries, buffers). It seems unlikely that express lane revenues will ever be sufficient to pay for the capital costs of the GP lanes as well as the tolled lanes.

Hence it is misleading to propose using the surplus revenues to pay for transit or equity-compensating income transfers when the gross revenue is insufficient to cover the full costs of the highway [56]. This makes the issue of exemptions (as well as multiclass service itself) an important factor in transportation financing, above and beyond the short-run economic efficiency impacts.

With the entire road priced – whether single- or multiclass – the revenues from efficient pricing can more than cover the long-run capital costs of the facility, and the surplus applied to capacity expansion of the highway, improved transit service, or equity transfers. Dividing revenues in thirds to highway capital, transit service, and equity transfers is referred to as the Small rule (Small [57]), as applied to whole-facility pricing. Capacity expansion of the highway will lower the congestion toll rates, by reducing congestion, and the decision to expand should be based on estimated benefits and costs, not on the amount of revenue, although toll revenues are a good indicator of WTP for highway investment.

Some Variations in Configuration and Operation

Numerous variations in design and operation have been tried, for purposes of making a restricted class of service work successfully, which is to say the service is neither underutilized nor overutilized, and customers are not complaining unduly. A few of the most notable variations are described below.

Multiclass Networks If the concept of multiclass service is expanded to the metropolitan area, two broad strategies arise:

1. All potentially congestible expressways offer HOT lanes along with GP lanes, and separation of the services is maintained at interchanges with other expressways.
2. Some portions of the expressway network are fully priced and others are not, in a way that allows users to reach the same destinations via tolled or untolled paths.

The first strategy requires substantial infrastructure to permit all choices of direction to be continuously separated (i.e., express lane users never have to sit in GP lane congestion). Potentially, the costs of such infrastructure may be sufficiently large that they might outweigh the benefits from reduced congestion, especially in places that are not highly congested. The second strategy avoids many of the problems of multiclass service, including the above infrastructure costs, but different locations would be differentially served. Motorists would still have a range of choice, but not as much as when all expressways offer both GP and express services.

Shoulder Used for Travel Lane During Peak By incorporating the shoulder as a travel lane on a temporary basis during peak periods, an increase in capacity can be achieved at low capital costs and an unknown increase in safety costs.

A complementary strategy is to restripe the road for a larger number of narrower lanes. The full 12-ft width is not required for speeds below 50 mph, so more lanes can be squeezed into the same roadway width [58]. Speed limits need to be lower, which actually helps to smooth traffic flows during peak flows but is awkward during off-peak periods.

Discounts for HOVs Most HOT lanes in the USA were previously lanes restricted to high-occupancy vehicles (HOVs), and the tolling strategy was intended to utilize the excess capacity resulting from the discrete nature of carpool lane eligibility criteria (two- vs three-passenger minimums). In these cases, the tendency is to permit the HOVs to continue to use the lanes free or at a discount.

The effectiveness of HOV segregation as a stimulus for reducing peak vehicle trips through higher vehicle occupancy varies with the context. Prearranged carpooling has been in secular decline for several decades, but casual or opportunistic or unscheduled carpooling (e.g., slug lines) have spontaneously appeared when the particular conditions are suitable. Because the HOV concept originated as a substitute for pricing, it could be argued that keeping the HOV preference is both redundant and a move away from efficiency by overencouraging ridesharing. The addition of the ability to pay to use the lanes can also reduce demand for casual carpools. The number of casual carpoolers decreased dramatically on the Katy Freeway once the carpool lanes were opened to paying SOVs. This may be due both to much better travel along the freeway (high time-value drivers no longer need occupants to gain faster travel) and to passengers switching to toll-paying SOV mode (while there is no increased incentive to attract new volunteer passengers).

Number of Priced Lanes

The greatest penalty for reconfiguring an expressway for multiclass service occurs when the priced service consists of a single lane, because of the limited capacity of a single segregated lane (see Loss of Capacity above on p. 10). Two lanes are better because they allow for some weaving and therefore higher flows per lane. This suggests that a minimum of four or five lanes in the same direction is a requirement for creating efficient multiclass service.

Evaluation of Pricing Alternatives

Although there has been a great deal of writing, both theoretical and empirically based, on the performance of HOT lanes, no single analysis has taken in the full welfare consequences of HOT lanes in comparison with the relevant alternatives. Lee [20] urges an evaluation framework based on welfare measures for applicable alternative uses of the same capacity. Sullivan and Burris [59] made pioneering efforts to differentiate average VOT for toll payers versus non-payers. Light [60] has constructed an abstract framework that addresses the optimal price of one service that is priced, the optimal capacity with an explicit distribution of travel times and no exemptions, and no demand elasticity; he uses a modification of the "two route problem." Agency evaluations tend to focus on utilization and revenue, and seldom attempt to perform benefit-cost analysis [61].

Counterfactual

Many studies that assert benefits of HOV or priced lanes are either vague about what they are being compared to or pick a counterfactual that may not be the economically efficient use of the infrastructure. HOV lanes are judged by the difference in speed between GP lanes and HOV lanes, rather than the change in travel speeds for all users resulting from the dedication of one or more lanes to restricted usage (see for example [14] with respect to HOV lanes).

Within the context of managed lanes and highway pricing, there is a range of alternatives that needs to be made explicit.

	Single-class service	Multiclass service
No tolls	All lanes merged into GP	HOV lanes separated
Express lanes tolled	All lanes pay the same toll	Express lanes pay toll
All lanes tolled		Both services priced but with different toll rates

Efficient Congestion Pricing

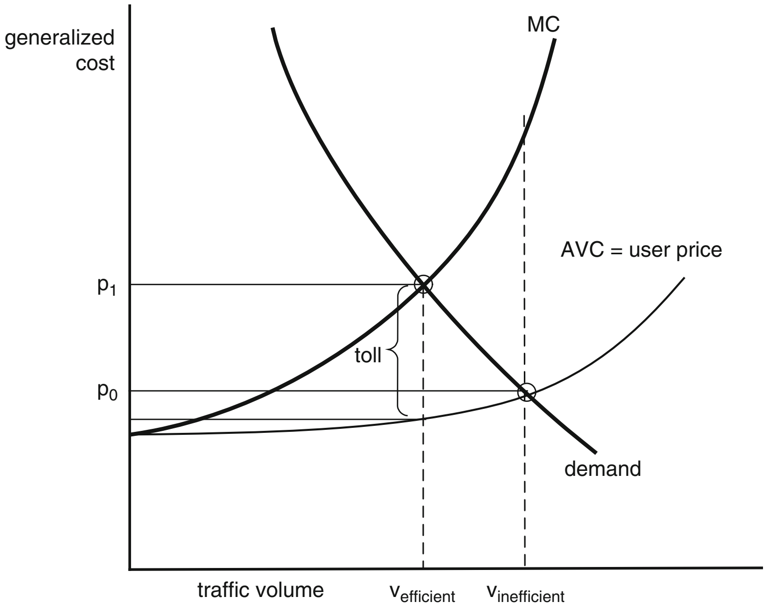
To evaluate the performance of priced lanes, it is necessary to review the basic theory of road pricing [62]. Without a compensating adjustment in price, the number of vehicles using a facility when congestion arises is greater than the number that would occur under economic efficiency. The efficiency condition is not met because some number of users value the travel (to themselves) at less than the incremental cost in delay time that they impose on the traffic stream as a whole. The reason they choose to travel anyway is because the user faces the average travel time cost rather than the marginal social time cost. The marginal is always larger than the average because the average rises with additional traffic.

Thus, without correction, the price to users on congested facilities is always lower than the efficient price. The efficient price does not eliminate all congestion, but it should be sufficiently high to permit free-flowing traffic (not necessarily free-flow speed). Maximum capacity is often defined [27] as a density of 45 vehicles per lane per mile. The efficient volume will be somewhere between the maximum flow and the volume that results in free-flow speed.

Any increase in price from the unpriced situation will result in a gain in efficiency, up to the point where traffic speed is higher than what the users would be willing to pay for to deter enough vehicles so that the speed can be maintained. Express lane managers typically seek a minimum speed of 45 mph during the peak period, which is probably close to efficient, depending upon the VOTs of the vehicles in the stream.

The Standard Road Pricing Model The basic theory of efficient pricing of roads applies to a single distinct segment of road or a bridge, with a given demand for vehicle trips (in vehicles per hour). Users bear a group of variable costs, the most significant of which is travel time. The facility has a flow capacity, depending upon the number lanes, curves and grades, and other attributes, measured in passenger car equivalents per hour. As vehicle flow increases and begins to approach capacity, travel speed declines at an accelerating rate, thereby increasing travel time (the reciprocal of speed). The relationships in their most simplified form are displayed in Fig. 2. The vertical axis is the generalized cost or generalized price, in which all variable costs including travel time are translated into dollars [63–65].

It is assumed that the user pays vehicle operating costs and accident risk, and correctly anticipates the average travel time during the period the user is on the facility. There are no user charges or excise taxes, at least of a nature that vary with the amount of use, or else



HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Figure 2
Basic highway congestion tolling

such fees exactly pay for the variable costs of maintaining the road for the particular user (pavement wear, law enforcement). There are no external costs that the user fails to pay for (e.g., pollution), and no external benefits outside of the demand curve. Fixed costs of the highway and the vehicle are not explicitly included in this analysis.

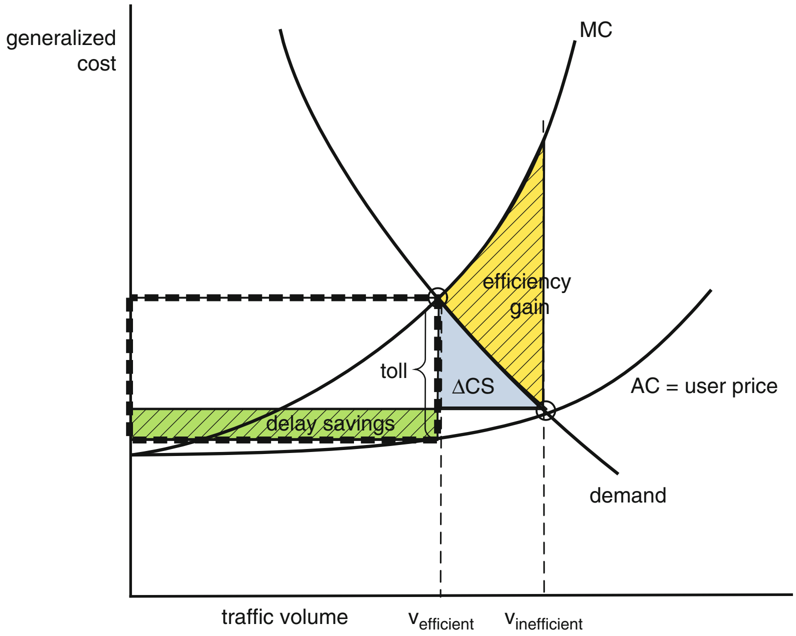
The result of these simplifications is that the price to the user is the same as average variable cost (AVC), whereas the true cost to society of the marginal vehicle is the marginal cost (MC), described heuristically as the AVC plus the delay created for others by the presence of the marginal vehicle. In that sense, each vehicle creates a negative externality in the form of delay to other vehicles. Usage of the facility occurs at an equilibrium volume higher than that which would maximize net benefits to society – that being where price = MC – from the availability of the facility.

The "overusage" is the volume between $V_{\text{inefficient}}$ and $V_{\text{efficient}}$, which results in an increase in costs represented by the area under the marginal cost curve between the two volumes, while the incremental benefit of this traffic is represented by the area under the demand curve (which is the same as the marginal social benefits curve) between the two volumes. The "toll" is a monetary charge levied on the vehicle, calculated to make the price to the user (per mile or per trip) equal to the marginal cost at the efficient equilibrium.

This inefficiency, indicated as a shaded triangular-like area in Fig. 3, consists of two components: an increase in consumer surplus (a positive benefit) from the value placed on the travel by the additional vehicles, above what they actually pay, and an increase in delay for "old" vehicles (those that would be on the facility if it were efficiently priced). In congested conditions the delay greatly exceeds the consumer surplus, the difference being the area of the inefficiency triangle (also called welfare loss). Mathematically,

$$\text{Inefficiency} = \text{delay savings} - \Delta\text{CS}$$

where the "inefficiency" is a loss or net disbenefit and ΔCS = incremental consumer surplus. The data needed to calculate these benefits consist of observed traffic volumes and speeds, estimated delay functions, other variable costs, and the generalized price to the user [66]. Throughput by service and time savings are not sufficient for evaluation, although they may be useful performance indicators. Kidd [67] compares numerous configurations of HOV, HOT, BRT, and light rail modes on performance measures emphasizing



HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Figure 3
Benefits of congestion pricing

throughput and speed, and recommends HOT lanes, in part for their potential BRT use. Poole [68] also emphasizes this combination.

The standard model also does not represent the possibility for hypercongestion, where capacity is lost by overloading the facility beyond its ideal maximum capacity. Peak pricing can prevent this from happening, even if the toll is well below the efficient price.

Dynamic Demand Schedule

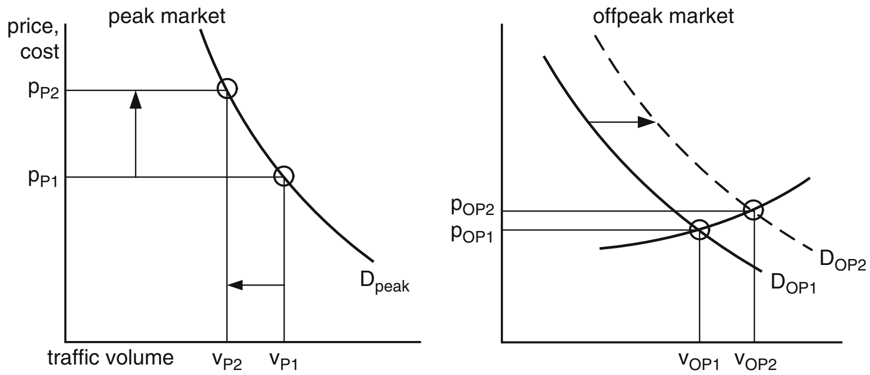
The above assumes a given static demand curve. If demand was always at the same level, the toll would be constant. The nature of the congestion problem is that demand is not constant, but instead has peaks and valleys. A benefit of the congestion toll then becomes the shifting of some users from peak periods into off-peak periods during which the price (toll) is lower or zero. In addition to the previously available user options of higher occupancy (carpool, bus transit), alternative mode (e.g., rail freight or transit), and forgoing the trip, the user can shift the time of the trip to an off-peak period. The more options available to the user, generally the higher (in absolute value) is the demand price elasticity, and hence the smaller the toll required to achieve an efficient equilibrium.

With multiple demand periods, there is an interdependency among them. Specifically, if the price is raised in any period (e.g., the peak), then the demand curve will shift outward in the other periods (e.g., the off-peak). This effect is shown in Fig. 4. An increase in peak price from p_{P1} to p_{P2} shifts the whole demand curve in the off-peak market from D_{OP1} to D_{OP2}; correspondingly, a price decrease causes demand in the related market to shift inward. The magnitude of the shift is measured as a cross-elasticity.

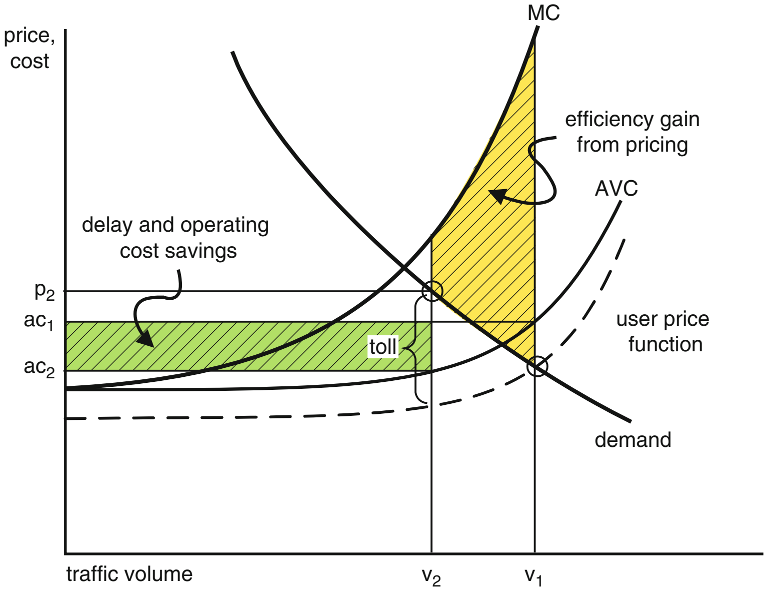
Peak versus off-peak periods illustrate the case of markets too closely related to be treated as separate markets. A similar situation of markets for close substitutes arises with multiclass service. There are a number of elasticities and cross-elasticities that result from changing prices in any one of the several closely related markets. New users will be attracted in each of the markets, at higher or lower occupancy levels, and others may be deterred from the same and other markets. For some simple purposes, the numbers of travelers, vehicles in the corridor, number of carpools, etc., can be approximated as remaining constant, but in general they will change. For example, when HOT lanes constructed and operated by private concessionaires promised guaranteed speed at all times, bus service was planned to use the expressways; previously the buses avoided the freeways due to unreliability of schedules [69]. How to value such changes makes evaluation of multiclass service difficult.

Nonoptimal Pricing

The standard model makes the convenient assumption that average cost and user price are the same, but in practice this only applies to travel time; for other costs (emissions, vehicle operation) and user charges (fuel taxes, flat tolls), any correspondence between cost and price is coincidental. Because for HOT lane policy it is necessary to analyze tolls separately from costs, a more realistic representation is needed. A general framework for evaluating transportation investment and pricing that recognizes the price function as a separate factor is shown in Fig. 5. In this case, price is shown as lying below AVC. Because both price and AVC include travel time, the two curves tend to have the same shape. In the



HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Figure 4
Peak and off-peak demand periods



HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Figure 5
Benefits of pricing when price is not efficient and not equal to AVC

example, the assumption is that external costs in the form of noise and air pollution emissions, plus agency costs for maintenance and administration, outweigh the variable user taxes paid by the vehicle. The intersection of the price function with the demand curve determines the initial volume v_1. The average cost (including social costs) at that volume is marked as ac_1. In reality, the price function might lie either above or below the AVC curve.

The user impact components from Fig. 3 still apply, but they are a little more complicated. In this example, the context is simply two different prices, neither of which is efficient. Although the optimum price is not achieved by the toll, there is an efficiency improvement from reducing volume from v_1 to v_2. The efficiency gain (reduction in the efficiency loss) now consists of three parts:

1. A reduction in excess delay (and operating costs to the extent they increase with congestion), indicated by the shaded rectangle
2. A loss of consumer surplus (not marked in the diagram), which is still a triangular area lying under the demand curve between v_2 and v_1
3. A reduction in net external costs that are above the price curve, represented by a rectangle (not marked) whose height is the difference between price and AVC at v_1 and whose length is v_1 minus v_2

Emissions changes can be estimated by modeling the conditions, which translate into speed change or "drive cycles" that alter the power load on the engine. Congested drive cycles are much different from constant speed cycles, and patterns on tolled lanes are different from GP lanes, but the differences and magnitudes are relatively small compared to time values. Estimating changes in consumer surplus (including the value of delay) is more difficult with multiclass service.

Network Effects

The congestion pricing model in Fig. 2 assumes implicitly that all facilities are optimally priced. The demand interactions considered so far have been between peak demand and off-peak demand. In a more generalized system or network, each link is a separate but interrelated market for travel. Each traveler seeks the minimum path through the network to his/her destination, considering operating costs and the value of time. In the face of an increase in money price on a particular link, the traveler may

- Continue to travel and pay the toll (high time value)
- Divert to another route or path (higher time cost for traveler, increase delay on alternative path)
- Shift to carpool or transit (perhaps offering better service because of reduced congestion)
- Shift time to off-peak (traveler worse off than without the toll, but not as bad as paying the toll)
- Change in location or origin or destination, long run or short run, or combine trips into tours (trip chaining)
- Forgo the trip

Each of these responses represents a valuation by the user of the available alternatives. The benefit of the price changes (whether positive or negative) is the sum of all changes in consumer surplus. Poole [68] has stated the welfare objective in the following way: ". . . a managed lane that charges all customers the market value of their trips (time plus reliability) maximizes the utility of those customers. The higher the total revenue such a lane produces, the greater the value of the transportation benefits it is providing. By contrast, the goal of maximizing the number of people per lane per hour implicitly values all those trips equally, and at a low value (the extra time and convenience of being in a carpool)."

Related Markets

The concept of a market, with supply and demand, and a resolution, whether a constrained or an unconstrained equilibrium, is an abstraction from the field of economics. Whether the concept can be applied to a given context, how it should be applied, and whether anything useful results depends upon the skill and insights of the analyst.

Inherent in the idea of a market is the understanding that markets are connected to each other: a change in one market affects all other markets, directly or indirectly. Such an understanding would be paralyzing (nothing can be concluded unless all markets are studied) without some principles for bounding the scope. Three ranges of scope can be considered:

1. Primary Market

   Responses taking place within the market are endogenous, and all relevant factors should be resolved; other markets may be incorporated indirectly via changes in consumer surplus, but are not recognized explicitly.
2. Related Markets

   Markets whose demand or costs are directly affected by the primary market may need to be included, such as peak and off-peak markets, as being not autonomous (e.g., peak and off-peak may share the same capacity).
3. Other Markets

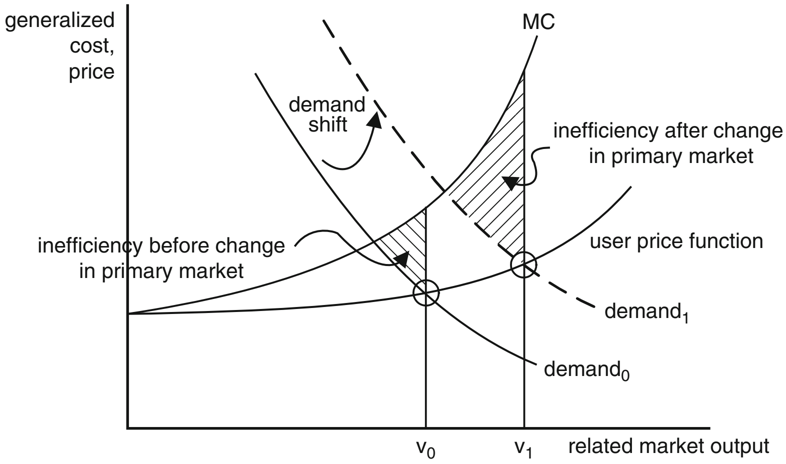
   Related markets that exhibit market failure (monopoly, constrained pricing) may be incorporated into the evaluation, to the extent that their inefficiencies are affected by pricing in the primary market.

All other markets not explicitly included are assumed to adjust efficiently to whatever changes occur in the primary market. When pricing a single facility, it is assumed that all other facilities are "other" markets, which adjust their prices and output (including investment in capacity) to remain efficient. Whether vehicle trips gained or lost in the primary market are the result of diversion, occupancy shifts, or deterred/induced travel is unimportant, or at least ignored. Shifts between peak and off-peak, however, are certainly relevant.

Multiclass service creates two markets that are closely related. Similarly to peak and off-peak demands, an increase in the price of the premium service causes an outward shift in demand for the standard service. If peak and off-peak are also distinguished, then there is a matrix of elasticities connecting the four options. An increase in the price of the premium service in the peak, for example, will increase demand for off-peak premium service and both standard service demand periods.

Evaluation of changes in related markets might be approached by treating the market in which the change occurs as the primary market, and changes in related markets as relevant. Benefits would be as described above for a single facility, plus changes in the magnitude of the efficiency loss in related markets. If all other markets, whether directly related or only indirectly, adjust efficiently, then nothing in those adjustments need be considered in evaluating the changes in the primary market. If, alternatively, related markets are inefficiently priced or constrained, then the change in inefficiency can be regarded as a benefit of the change in the primary market [70].

An increase in price in the express lane market, for example, would cause an increase in demand in the substitute GP market, as shown in Fig. 6. If the related



HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Figure 6
Change in inefficiency in related market

market is priced inefficiently, and the deviation between price and marginal cost is increased by higher output, then the inefficiency is worsened by the price change in the primary market. The change in efficiency (a negative number, in this case) could be added to the benefits of the price change.

Some Lanes Priced

If the theory shown in Fig. 2 is applied to one or several lanes of an urban expressway, but not to all lanes, an improvement in efficiency can be obtained even if the pricing is not optimal, as in Fig. 5. Vehicles tolled off the restricted lanes may take any of the options listed above, including consolidation into more person trips in the remaining vehicles. At least some of the vehicles, however, will choose to not to pay the price but will use the remaining GP lanes. This demand shift in the closely related market, as in Fig. 4, would not be accounted for in the benefits (disbenefits) from pricing.

One could assume that all other markets are efficient and make optimum adjustments to price changes in related markets, but this assumption is clearly violated in practice. The price is fixed at an inefficient level (unlike the stopped clock, it is never correct), and the capacity cannot be varied either. Inevitably, the inefficiency that is (in this case) worsened in the related market needs to be addressed.

If a portion of the traffic is separated from the rest and charged a price to keep it flowing at LOS C, the price on those lanes is higher than without the pricing, and the delay is greater on the GP lanes if the total travel remains the same (more demand is imposed on less capacity). Therefore the average generalized price is greater and the result is that the total traffic volume will be less. Vehicle occupancy is raised by the pricing (and not by congestion), but this reduction in vehicles will less than offset the additional demand directed at the GP lanes. Under a reasonable range of conditions, then, time savings on the express lanes may be more than the increased delay on the GP lanes (because the average user price is higher and some efficiency is gained) or may be less than the delay on the GP lanes because not enough vehicles have been deterred to offset the capacity loss from separation. This assumes, of course, that there is no increase in the total capacity of the facility.

All Lanes Priced Uniformly (Single Service)

The prototypical condition imagined in the standard theory is steady flow on a single uninterrupted facility. In this abstract market, there is a single demand curve and single cost curve. Cost is determined primarily by the relationship between traffic volume and delay. Marginal cost can be derived from the average flow relationship (empirically or analytically) and the optimal toll calculated and imposed. The gains from this congestion pricing is a reduction in delay for the users choosing to the pay the toll, minus the loss of consumer surplus suffered by the deterred vehicles, shown previously in Fig. 3. The number of travelers may stay the same (or even grow) but some of them have been inconvenienced by having to share the same vehicle.

This simplified abstract model does not say anything about the distribution of travel time values among the vehicles along the demand curve. As the price goes up, some users are deterred by the sum of time and money costs, but the trade-off between them will not necessarily be the same at any point along the demand curve. Moving up the demand curve, it is implicitly assumed that the money price is increasing while the delay portion of generalized price is decreasing.

Thus the standard model is robust for the mechanism it uses (a single price for all users), but does not acknowledge that conditions may be considerably more complicated.

Both Services Priced Efficiently

The fundamental question in regard to multiclass service is whether there are market segments with the demand for the service that are sufficiently distinct as to warrant separate tailored services along with the additional costs that differentiation entails. In the case of highway travel, the question is what distribution of travel time values among the vehicles potentially demanding service justifies differentiating the demand into two or more classes and charging different prices. Clearly, charging a price for both services is a gain in benefits over charging for only one service, but are the optimal prices charged for the two classes different enough to justify the loss in capacity?

There seems to be little analysis or evidence on this problem. Many analysts assert that the distribution of

VOTs is wide, but how wide in truth and what width is needed to take the extra step of multiclass service is not yet solved.

To evaluate a multiclass service is to address the problem of allocating a fixed amount of capacity to two classes of users, each with its own demand (reflecting differing preferences for travel time, occupancy, and other attributes) and pricing structure. The resulting equilibrium is then compared to the base case.

First-best evaluation is a comparison of alternatives all of whom are efficiently designed and priced. Second-best evaluation occurs when at least one constraint is imposed on at least one of the alternatives, and the net benefits of one alternative versus another are estimated.

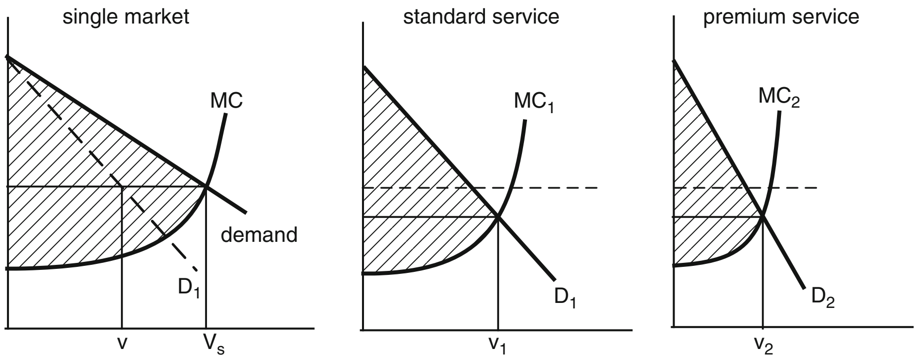
A first-best evaluation of multiclass highway service would compare an efficiently priced single facility to two efficient service alternatives. Total capacity would be fixed, the two services would be designed with demand attributes that maximized the net benefits to users, including pricing at marginal cost. The conclusion as to whether two-class service was better than a single class would derive from the net social benefits (user benefits minus implementation costs) for each alternative.

A hypothetical example is shown in Fig. 7. A single-class service is portrayed on the left, and a two-class service in the two markets on the right. Each market is resolved efficiently at p = MC, and the net benefits of each – the area above the MC curve and below demand, up to the equilibrium volume – are shaded for each service. In the single-class service, a dashed demand line indicates the separation of demand into those who will choose each one when it is offered.

Once the two types of service are offered, each has a separate demand curve (related in that a change in price in one of the markets will shift demand in the other; thus each demand curve is based on a given price in the other market). By separating the two market segments, the attributes of each service can be tailored to its customers. The two demand curves are not necessarily a partition of the single-service demand: new users may be induced by the existence of, especially, the premium service. With differing prices, movements along the demand curves may result in greater or lesser total volume, i.e., the sum of v_1 and v_2 may be larger or smaller than V.

The problem is greatly complicated by the need to distinguish different values of time. The standard efficient pricing diagram assumes that time and monetary costs can be combined on a single scale, implying that the ratio of time to dollars – the value of travel time – is fixed. Thus the two-dimensional diagrams in Fig. 7 are only slices through a three-dimensional space in which demand is represented as a surface that is a function of both time and money. If a third dimension to the diagram is imagined, the standard service is a slice through the low-time-value end of the diagram while the premium service is a slice at the high VOT end. The single market is a slice through the middle, or at average VOT.

In the particular representation in Fig. 7, the generalized price to the user in both markets is below what it was in the combined market because the money price



HOT Lanes/Value Pricing: Planning and Evaluation of Multiclass Service. Figure 7
Benefits of two-class service relative to single service

is lower and the time cost is less highly valued. Hence, total net benefits are higher for multiclass than for single service. For the premium service, the price is lower because the higher money price is more than offset by the faster travel. For the GP lanes, travel times are worse than in the combined market, but are valued less highly than the reduced price. These results are not true in general, but occur only if the service characteristics and prices very closely match the demands of each group. Otherwise, the generalized prices combining toll and travel time will be close to the single market price or above it, and the multiclass service will offer no advantage over the efficiently priced single-class service. Keep in mind that all services – combined or separated – are priced efficiently in this comparison.

The provision of two classes of service creates a net benefit if the sum of the benefits for standard and premium services exceeds the benefit of the single service by enough to offset the costs of providing two services. If all travelers are homogeneous in their preferences, then there are no benefits from multiclass service because all users want the same thing. This fact and several other aspects of the problem are addressed by Small and Yan [71]. Thus the two market segments must have distinctly different demand curves, and the service classes must have attributes that separate the demand segments. The benefit of multiple services relative to a single service depends upon how different the market segments are, and how well the service designs appeal to and satisfy those differences. The theoretical basis for this evaluation has been most completely developed by Small et al. [45], but they did not incorporate any penalties for partitioning the facility into service classes.

Empirically, attempting to measure the benefits by valuing the attributes (such as travel time differences) is not likely to be useful, because less easily observed differences are too hard to measure or observe (e.g., the cost of schedule shift, or coordinating a carpool). Revealed preference can show demand quantities and generalized price at equilibrium, but the remainder of the demand curve is a problem, as are the cross-elasticities.

A paper from the operations research literature [72] describes a model for allocating customers to two classes of service according to their willingness to pay for service time. The model determines the capacity given to each service and the profit-maximizing prices charged. Possibly a model of this form could fit the multiclass service optimization problem.

Equity

Any form of pricing of highways raises numerous claims of inequitability. The concept of HOT lanes generates additional objections beyond those that occur in the context of “simple” (single class) tolling and congestion pricing because there is a side-by-side comparison. People observe that, among toll payers, the burden of payment declines as a share of income as income rises, so pricing is regressive. Others observe that middle and higher income users are the majority among payers, so they are getting the benefits, and that is regressive. These both reveal an incomplete understanding of distributional equity.

Lexus Lanes Vertical equity impacts (impacts on the distribution of income) can be addressed objectively and quantitatively, although the data are coarse at best. Among travelers, peak highway users have higher average incomes than both transit users and off-peak users. Where express lanes are available, income levels of toll payers are usually higher than non-payers (Sullivan and Burris [59]; FHWA [73]). As the income of toll payers decreases, vehicle occupancy usually goes up. Hence the accusation that express toll lanes disproportionately benefit higher income users. The counter anecdote to the Lexus lane barb is the observation that working parents use the express lanes to pick up their children from daycare. Despite the expressions of concern, equity discomfort tends to subside after initial reaction to proposed HOT lane conversions, and has not so far been an impediment to implementation [55].

Unlike tax instruments, both the costs and benefits of paying for a market good are borne directly by the buyer. The buyer weighs the costs and benefits to himself/herself and makes a choice. If the consumer chooses to buy, then the benefits outweigh the cost, possibly by a large amount but a magnitude not apparent to the outside observer. If there are no external costs and no subsidies, then the social benefit of the good or service is at least as great as its costs, and neither consumers nor taxpayers are disadvantaged.

The existence of the good may divert consumption from other goods, but there are no inherent equity impacts, i.e., income is not being transferred without compensation. The above notwithstanding, "some argue the HOT lanes unfairly favor those wealthy enough to afford high tolls" [74].

This leaves the only complaint being that the underlying distribution of income is unfair. The solution, then, is a "Robin Hood" income transfer that rearranges income solely according to income, not according to road usage or any other criterion. The poor will then have the wherewithal to purchase peak SOV travel, or use the income for some other purpose.

Perceptions of equity are difficult to address in rational terms. Resentment against those who choose to buy their way out of congestion is not really any different from resentment against those who buy luxury yachts and large mansions, except that the cost of buying out of congestion is far less.

Earned Credits In an effort to improve the public acceptability of tolls, it has been proposed that drivers earn credits by using the unpriced lanes. The credits could then be applied to use of the express lanes [75]. Other proposals have been offered [76]. No projects have so far been implemented.

Generally credits are earned for purchasing something, as opposed to *not* purchasing. Thus credits would be awarded for paying the tolls for some period of time or amount, and a bonus paid to the user. The purpose is to create brand loyalty, which is not relevant here.

Hence the purpose is based on equity concerns. Sometimes it is vertical equity that is of concern, in which case the credits should be awarded to all persons or households solely on the basis of income, and unrelated to whether the recipients actually drive in the peak period or would like to. Rewarding drivers for using the GP lanes does not seem to accomplish any equity goal and is perverse from an efficiency standpoint (it encourages peak usage).

However credits are awarded, they constitute a financial transfer that reduces the net revenues from the tolls. Other losses of revenue include collection costs, operating subsidies for transit, and matched reductions in unrelated taxes such as property taxes. If the proceeds are rebated to users of the facility *because of usage*, they have a negative efficiency impact as well as an equity impact.

Land Use Impacts The rent gradient model of urban form relates the rate of decline in land value with distance from a point of central access to the unit cost of transportation. In this model, a reduction in the price of transportation to the user tends to flatten the rent gradient, causing land use activities to spread out at lower densities. Other things being equal, an increase in capacity reduces the travel time from delay and leads to dispersal of economic activity. Correspondingly, an increase in prices – e.g., through higher prices to users – should result in more compact development. While whole-facility pricing shifts the relative spatial preferences of households, on average the price is higher with efficient pricing. Thus the standard congestion pricing model applied to an urban network will tend to steepen the rent gradient.

HOT lane pricing does not lead to unambiguous results when interpreted via the rent model. While the explicit money price is greater on the tolled lanes, delay is greatly reduced; the net result, however, is an increase in price – for the same capacity. HOT lanes have never, however, been taken from existing capacity. The one attempt to do so – the infamous Santa Monica freeway diamond lane in Los Angeles – was a political disaster sufficiently painful to ensure that removing existing lanes has not been attempted again, and extensive outreach is undertaken before any kind of reserved lane is implemented [77]. Even if the amount of capacity added is less than what would be added by all GP lanes, the result is nonetheless to increase capacity and lower the cost of travel, spreading out the urban area.

Future Developments

Although the mileage of highways in "pure" HOV facilities vastly exceeds the tolled mileage, there are few plans to add HOV-only capacity; treatments being considered include general capacity expansion and tolling [16].

DKS Associates [10] describes activities in ten metro areas, most of which do not have congestion pricing currently, and some that do are not included in the survey. The report also describes how HOT lanes can be incorporated into the regional planning process, and offers lessons learned regarding implementation from each of the areas.

Demonstration Effect

For at least some analysts, the HOT lane was a way to leverage the perception that HOV lanes were underutilized into a demonstration that pricing works. In this scenario, the mechanism of pricing could substitute for the awkwardness of verifying occupancy, and the occupancy requirement could then be dropped because the pricing would have the same effect. With pricing established as a management tool, all lanes could be priced and the physical separation no longer necessary. All links of the network (or the arterial network) could be readily tolled, and traffic flow greatly improved. With no other evidence of political willingness to try congestion pricing, HOT lanes was the only game.

In this way, the expansion of the concept could encompass all roads, or at least urban highways. Fielding and Klein [78] were early promoters of the spread of congestion pricing by imitation. The next level in this scenario is the congestion pricing of whole facilities in a strategic pattern. The planned pricing of the SR 520 bridge in the Seattle area will be the first example of such whole-facility congestion pricing in the USA, if and when it occurs.

Evolution of Multiclass Networks

Poole [79] has described a scenario in which single-lane HOT lanes phase out the HOV exemption or discount, because the exemptions prevent the revenue streams from being large enough to cover capital costs. These more-or-less fully priced one-lane highways are then successful enough to construct the ramps and flyovers that permit the express service to be continuous when one freeway intersects another. With this success as a model, the one-lane priced highways can be expanded to two lanes in each direction, resulting in a reasonable enough level of capacity and associated throughput to justify the separation.

An alternative scenario has emerged. Travelers are greatly enamored of the freedom to choose between tolled and untolled lanes, and HOT lanes are being extended into separated networks. The expense of the infrastructure for maintaining separation at interchanges has not deterred investment in HOT lanes. Although the benefits in time savings may not warrant the construction costs in many circumstances, such calculations do not seem to have been done and there is not an obvious political market for them.

From HOV to BRT The San Francisco Bay Area plans to evolve from HOV lanes to a single-lane HOT lane network with physical separation at interchanges, and from there to a tolled network that is heavily used by express buses [80]. The region plans to add capacity by utilizing shoulders and relaxing freeway design standards. HOVs will remain free or discounted. Realization of this vision still implies a continuing primary contribution to transportation funding from the federal government.

From HOV to HOT Texas seems to be moving in the direction of two-class service with express lanes having at least two lanes (generally permanent) in each direction on both service classes [41]. This will entail considerable expenditures for physically separated infrastructure, yet HOVs appear to remain free and BRT does not carry a large share of passengers. The Texas DOT has produced plans for public–private partnerships to build and finance major multimodal intercity corridors, but there has been some political backlash from the legislature.

From HOV to Full Facility Tolling The Puget Sound region has an extensive existing HOV network, some of which the region expects to convert to HOT along the way to creating a network of fully tolled expressways with no differentiated classes of service and no exemptions or discounts [81]. While this end state is the most likely to be sustainable in the long run, the path for moving to there from HOVs is not presently spelled out.

From HOV to HOT and Truck Lanes The Reason Foundation has proposed an evolutionary path that leads from express lanes to multiclass service in which all services are priced and truck-only lanes are a candidate service [79]. Poole claims that these characteristics will emerge once federal subsidies are limited to highways of national interest and local regions take responsibility for most of the transportation infrastructure, including its funding. The resolution will be a system paid for primarily by users.

Inevitably there are limits to the expansion of simple (i.e., unpriced) HOT lanes in metropolitan areas. Swisher et al. [82] portray various longitudinal scenarios showing how express lane viability (i.e., neither too much nor too little traffic) can be retained by pricing selected users, including carpools. The ideal timing for introducing occupancy and pricing policy changes depends upon the growth of demand in the corridor.

At present, the prospects of universal road charges (VMT charges graduated by congestion and other costs) appear uncertain and not imminent, although there are some indications that the eroding contribution of fuel excise taxes and the lack of other revenue sources are improving the climate for tolling [83]. It may be that the inefficiencies of congestion will be tolerated in many locations, as well as the inefficiencies of multiclass services.

Public–Private Partnerships The USA is beginning to explore various arrangements that allow for more flexible management of the highway asset procurement process. Design-build construction and its extension into operation and financing have been used in some instances to create transportation infrastructure. Availability payments can allow the public sector owner to postpone payments to the builder until the facility is in operation.

If a project can cover its operating costs from tolls, then the possibility exists for letting a private firm operate the road and collect the tolls, potentially also setting the toll rates. One of a planned series of HOT lanes in the Washington, DC area is being constructed via a public–private partnership between the State of Virginia, the Federal Highway Administration, and Transurban-Fluor [84, 85]. The first modern era US toll lane facility was financed and constructed by a private firm, although it was eventually necessary for the State of California to buy out the private firm. SR-91 opened in 1995 and was repurchased in 2002 by the Orange County Transportation Authority [86]. The logic of letting some vehicles pay extra for speed creates an opportunity for letting the private sector front the funds to construct the additional lanes. Planned toll lanes in Sacramento may be funded in this way [87].

Using General Revenues The alternative to toll revenues for financing is general fund revenues. Many states and localities have passed referenda to pay for transportation with sales taxes. The federal government has transferred $30 billion into the Highway Trust Fund to keep it solvent. Perhaps because most voters see themselves as highway users, they favor spreading the costs over more payers than just users. This scenario maintains strong controls at the federal level over the raising of revenues and how they are spent, with Congress making decisions based on political balancing.

Whether the path ahead lies through extended multiclass service, private equity, and political devolution, or major general fund support for transportation is subject to conjecture and, we hope, debate.

Note

Information contained in this chapter consists of the judgment of the authors and does not necessarily represent the policy of the Volpe Center or the U.S. DOT.

Bibliography

1. Federal highway administration (2008) Managed lanes: a primer. US DOT/FHWA, Washington, DC. http://ops.fhwa.dot.gov/freewaymgmt/pubs.htm#mngd_lns_hov. Accessed 27 June 2011
2. Port authority of New York and New Jersey (2010) Bridges and tunnels, tolls, PANYNJ. http://www.panynj.gov/bridges-tunnels/tolls.html. Accessed Aug 2010
3. Small KA, Gomez-Ibanez JA (1998) Road pricing for congestion management: the transition from theory to policy. In: Button KJ, Verhoef ET (eds) Road pricing, traffic congestion and the environment: issues of efficiency and social feasibility. Edward Elgar, Cheltenham, pp 213–246
4. Collier T, Goodin G (2004) Managed lanes: a cross-cutting study. Prepared for Federal Highway Administration, Texas Transportation Institute, College Station
5. Sallach M, Vu P, Zhan YJ (2010) Corridor level analysis of potential Managed lanes investment on I-75 South in Atlanta, Georgia. Paper for TRB, Georgia State Road and Tollway Authority, Atlanta
6. Mirshahi M, Obenberger J et al. (2007) Active traffic management: the next step in congestion management. Prepared for DOT/FHWA, US, Washington, DC. http://international.fhwa.dot.gov/pubs/pl07012/. Accessed 27 June 2011
7. Federal highway administration (2010) Value pricing projects involving tolls. US DOT/FHWA. http://ops.fhwa.dot.gov/tolling_pricing/value_pricing/projects/involving_tolls/priced_lanes/index.htm. Accessed Aug 2010

8. Martin PT, Vladisavljevic I, Yusufzyanova D (2007) The I-15 Express lanes evaluation. Prepared for Utah Department of Transportation, Salt Lake City, University of Utah. http://www.udot.utah.gov/expresslanes/dld/U%20of%20U%20Express%20Lanes%20Evaluation%20November%202007.pdf. Accessed 27 June 2011
9. Liu X, Zhang G, Wu YJ, Wang Y (2009) Analyzing system performance for Washington state route 167 high occupancy toll (HOT) operations. Prepared for TRB, University of Washington, Seattle
10. DKS Associates (2009) A domestic scan of congestion pricing and managed lanes. Prepared for Administration, DKS Associates in association with PBSJ and Jack Faucett Associates, Federal Highway, Oakland
11. Federal Highway Administration (2008) Technologies that enable congestion pricing: a primer, US DOT/FHWA, Washington, DC. http://ops.fhwa.dot.gov/tolling_pricing/value_pricing/publications.htm. Accessed 27 June 2011
12. Goodall N, Smith BL (2009) What drives single occupant traveler decisions in HOT lanes? : an investigation using archived traffic and tolling data from the MnPASS Express Lanes. Paper for TRB, University of Virginia, Charlottesville
13. Brinckerhof P (2003) A guide for HOT lane development. Prepared for Federal Highway Administration, Parsons Brinkerhoff with Texas Transportation Institute in partnership with US DOT/FHWA, Washington, DC
14. PB Study Team (2002) HOV performance program: evaluation report, Los Angeles county metropolitan transportation authority, Los Angeles
15. 407 ETR (2010) ETR 407 express toll route, 407 ETR Concession Company Limited, Toronto. http://www5.407etr.com. Accessed 27 June 2011
16. Fuhs C, Obenberger J (2002) HOV facility development: a review of national trends. Prepared for TRB, Parsons Brinkerhoff, Houston
17. Burris MW, Lipnicky K (2009) HOV or general purpose lanes? Public Works Manag Policy 14(2):130–147
18. Poole Jr RW (2008) The HOT topic: enforcement in the fast lanes. Tolltrans, pp 44–49
19. Poole Jr RW, Kenneth Orski C (1999) Building a case for hot lanes: a new approach to reducing urban highway congestion. Reason Public Policy Institute, Los Angeles
20. Lee DB Jr (2008) Toward the evaluation of value pricing. Transport Res Record 2079:71–78
21. Chang M, Wiegmann J, Smith A, Bilotto C (2008) A review of HOV lane performance and policy options in the United States. Prepared for FHWA, Booz Allen & Hamilton and HNTB, McClean
22. Booz Allen Hamilton, and HNTB (2008) A compendium of existing HOV lane facilities in the United States. Prepared for Federal Highway Administration, BAH and HNTB, McClean
23. Cervero R (1999) Reviving HOV lanes. Transport Quart 53(4):67–81
24. Sas M, Carison S, Kim E, Quant M (2007) Considerations for high occupancy vehicle (HOV) to high occupancy toll (HOT) lane conversions guidebook. Prepared for Federal Highway Administration, HNTB/Booz Allen Hamilton, McClean
25. Plotz J, Konduri KC, Pendyala RM (2009) To what extent can HOV lanes reduce vehicle trips and congestion?: An exploratory analysis using national statistics on HOV mode use. Paper for TRB, Arizona State University, Tempe
26. Supernak J, Kaschade C, Steffey D (2003) Dynamic pricing on I-15 in San Diego: impact on travel time and its reliability. Transport Res Record 1839:45–54
27. Transportation Research Board (2000) Highway capacity manual: HCM 2000. TRB, Washington, DC
28. Barboza T (2009) O.C. Toll road price drops to attract commuters. Los Angeles times, 27 Mar 2009
29. Arizona Republic (2007) More Arizona drivers disobeying HOV rules. Arizona Republic, 31 Oct 2007
30. Jang K, Chung K, Ragland DR, Chan CY (2009) Safety performance of high-occupancy-vehicle facilities. Transport Res Record 2099:132–140
31. Cassidy MJ, Daganzo CF, Jang K, Chung K (2006) Empirical reassessment of traffic of traffic operations: freeway Bottlenecks and the case for HOV lanes. Institute of Transportation Studies, Berkeley
32. Anderson T (2007) Car-Pool lanes may be switched to toll lanes. Los Angeles Daily Breeze, 30 Nov 2007
33. AASHTO (2003) A manual of user benefit analysis for highways. American Association of State Highway and Transportation Officials, Washington, DC
34. Cassidy MJ, Bertini RL (1999) Some traffic features at freeway Bottlenecks. Transport Res (B) 33:25–42
35. Jung S, Wunderlich K (2008) Roadway network productivity assessment: system-wide analysis under variant travel demand. Prepared for Federal Highway Administration, Noblis, Falls Church. http://ops.fhwa.dot.gov/tolling_pricing/value_pricing/publications.htm. Accessed 27 June 2011
36. Booz Allen Hamilton, HNTB (2008) Policy options evaluation tool for managed lanes (POET-ML) users guide and methodology description. Prepared for Federal highway Administration, BAH and HNTB, McLean
37. Yan J, Small KA, Sullivan EC (2002) Choice models of route, occupancy, and time of day, with value-priced tolls. Transport Res Record 1812:69–77
38. Burris MW, Goel R (2009) Mode choice due to HOV to HOT conversions. Paper for TRB, Texas A&M University, College Station
39. Zmud J, Bradley M, Douma F, Simek C (2007) Attitudes and willingness to pay for tolled facilities: a panel survey evaluation. Transport Res Record 1996:58–65
40. Burris MW, Ungemah DH, Mahlawat M, Pannu MS (2009) Investigating the impact of tolls on high-occupancy-vehicle lanes using managed lanes. Transport Res Record 2099:113–122
41. Sullivan E (2000) Evaluating the impacts of the SR 91 variable toll express lane facility. Prepared for California Department of Transportation, Cal Poly State University, San Luis Obispo. http://ceenve3.civeng.calpoly.edu/sullivan/SR91/final_rpt/finalrep_full.pdf. Accessed 27 June 2011
42. Regan ED (2010) Pricing trade-offs: revenue vs. traffic management. Presentation to the National Road Pricing Conference,

Houston TX, Wilbur Smith Associates, Atlanta. http://tti.tamu.edu/conferences/nrp10/program/. Accessed 27 June 2011

43. MacGregor ME (2010) How revenue and traffic management are combined in the North Tarrant Expressway (NTE) managed lanes project in Fort Worth, TX. Presentation to the National Road Pricing Conference, Houston TX, Texas Department of Transportation, Dallas. http://tti.tamu.edu/conferences/nrp10/program/. Accessed 27 June 2011
44. Patil S, Burris M, Shaw D (2009) A comparison of survey designs for stated choice models: an application to commuting on managed lanes. Prepared for TRB, Texas A&M University, College Station
45. Small KA, Winston C, Yan J (2006) Differentiated road pricing, express lanes, and carpools: exploiting heterogeneous preferences in policy design. Brookings – Wharton Papers on Urban Affairs, Washington, DC, pp 53–96
46. Weiss E (2008) Will drivers pay to hurry up and wait? Washington Post, 25 Feb 2008
47. Brownstone K, Small KA (2005) Valuing time and reliability: assessing the evidence from road pricing demonstrations. Transport Res (A) 39:279–293
48. Joint Transport Research Centre (2009) Improving reliability on surface transport networks, OECD/International Transport Forum, Paris, France
49. Lam CT, Small KA (2001) The value of time and reliability: measurement from a value pricing experiment. Transport Res (E) 37:231–251
50. Patil S, Burris M, Shaw D, Concas S (2010) The value of travel time savings on managed lanes for urgent situations. Prepared for TRB, Texas A&M University, College Station
51. Cho Y, Goel R, Gupta P, Bogonko G, Burris M (2011) What are I-394 HOT lane drivers paying for?. Prepared for TRB, Texas A&M University, College Station
52. Poole Jr RW (2006) Managed lanes: what should we maximize? Surface Transport Innovations, Dec 2006
53. Poole Jr RW (2006) Transportation policy review: competing models of HOT lanes. Public Works Financing, July–Aug 2006, pp 25–26
54. Stone R (2010) High tolls lure drivers to I-95's pay lanes, Miami Herald, 31 Mar 2010
55. Bhatt K, Higgins T, Berg JT (2008) Value pricing pilot program: lessons learned. Prepared for Federal Highway Administration, K.T. Analytics and Cambridge Systematics, Washington, DC. http://ops.fhwa.dot.gov/tolling_pricing/value_pricing/publications.htm. Accessed 27 June 2011
56. Cuff D (2009) MTC approves plan for toll lanes, Contra Costa Times, 23 Apr 2009
57. Small KA (1992) Using the revenues from congestion pricing. Transportation 19(4):313–333
58. Ng CF, Small KA (2008) Tradeoffs among free-flow speeds, capacity, cost, and environmental footprint in highway design. University of California, Irvine
59. Sullivan E, Burris M (2006) Benefit-cost analysis of variable pricing projects: SR-91 express lanes. J Transport Engin 132(3):191–198
60. Light T (2009) Optimal highway design and user welfare under value pricing. J Urban Econ 66(2):116–124
61. Loudon WR, Synn J, Miller H (2010) Consideration of congestion pricing and managed lanes in the metropolitan transportation planning process. Prepared for TRB, DKS Associates, Oakland
62. Federal Highway Administration (2008) Economics: pricing, demand, and economic efficiency, a primer. US DOT/FHWA, Washington, DC
63. Mohring H (1976) Transportation economics. Ballinger, Cambridge, MA
64. Vickrey WS (1963) Pricing in urban and suburban transport. Am Econ Rev 53:452–465
65. Walters AA (1968) The economics of road user charges. International Bank for Reconstruction and Development, Washington, DC
66. Mohring H (2006) Congested roads: an economic analysis with twin cities' illustrations. In: Roth G (ed) Street smart. Transaction Publishers for the Independent Institute, New Brunswick, pp 141–170
67. Kidd B (2006) Multimodal optimization of urban freeway corridors. Prepared for Arizona Department of Transportation, Lee Engineering, Phoenix
68. Poole Jr RW (2010) Managed lanes: networks vs. individual facilities. Prepared for TRB, Reason Foundation, Plantation
69. Craig T (2007) Deals clinched on HOT lanes, Washington Post, 21 Dec 2007
70. Verhoef ET, Nijkamp P, Rietveld P (1996) Second-best congestion pricing. J Urban Econ 40(3):279–302
71. Small KA, Yan J (2001) The value of "value pricing" of roads: second-best pricing and product differentiation. J Urban Econ 49:310–336
72. Celik S, Maglaras C (2005) Dynamic pricing and leadtime quotation for a multi-class make-to-order queue. Columbia University Business School, New York
73. Federal Highway Administration (2008) Income-based equity impacts of congestion pricing: a primer, US DOT/FHWA Washington, DC. http://ops.fhwa.dot.gov/tolling_pricing/value_pricing/publications.htm. Accessed 27 June 2011
74. Cassidy WB (2010) Congestion pricing advances in virginia. J Commerce, May 27. http://www.joc.com/government-regulation/congestion-pricing-advances-northern-virginia
75. Buckeye KR, DeCorla-Souza P, Lari A, Aultman S (2009) User perceptions of fee lane concepts in Minnesota. Minnesota Department of Transportation, St. Paul
76. Gulipalli P, Kalmanje S, Kockelman K (2008) Credit-based congestion pricing: expert expectations and guidelines for application. J Transport Res Forum 47(2):5–20
77. Billheimer JW (1978) The Santa Monica freeway diamond lanes: evaluation overview. Transport Res Record 663:8–16
78. Fielding GJ, Klein DB (1993) "High occupancy/toll lanes: phasing in congestion pricing a lane at a time", Reason Foundation Policy Study 170. Reason Foundation, Santa Barbara
79. Poole Jr RW (2010) Pricing and Fiscal implications of different types of HOT Lanes. Presentation to the National Road

Pricing Conference, Houston TX, Reason Foundation, Santa Monica, CA. http://tti.tamu.edu/conferences/nrp10/program/. Accessed 27 June 2011

80. Klein L (2010) 2010 bay area express lane network. Presentation to the National Road Pricing Conference, Metropolitan Transportation Commission, Houston, Oakland. http://tti.tamu.edu/conferences/nrp10/program/. Accessed 27 June 2011
81. Stone C (2010) SR 520 Lake Washington Project: Seattle's urban partnership agreement. Presentation to the National Road Pricing Conference, Houston, Washington State Department of Transportation, Olympia. http://tti.tamu.edu/conferences/nrp10/program/. Accessed 27 June 2011
82. Swisher M, Eisele WL, Ungemah D, and Goodin GD (2003) Lifecycle graphical representation of Managed HOV lane evolution. In: 11th International HOV conference, Seattle
83. Copeland L (2010) States turn to tolls to fund roads. USA Today, 2 Aug 2010
84. Adams T (2010)Toll concession model for the I-495 beltway HOT lanes. Presentation to the national road pricing conference, Transurban, Houston, Washington, DC 2–4 June 2010. http://tti.tamu.edu/conferences/nrp10/program/). Accessed 27 June 2011
85. Weiss EM (2008) Money spent vs. time saved debated for HOT lanes. Washington Post, Aug 25. http://www.washingtonpost.com/wp-dyn/content/article/2008/08/24/AR2008082402559.html
86. Miles M (2007) Tolling & value pricing: the California experience. In: ITE – Technical Conference 2007, California Department of Transportation, San Diego
87. Turner M (2008) Officials look to toll lanes to solve traffic, funding issues. Sacramento Bus J, Feb 22. http://www.bizjournals.com/sacramento/stories/2008/02/25/story1.html

Human Bacterial Diseases from Ocean

Darrell Jay Grimes[1], Lisa W. Plano[2], Okechukwu Ekenna[3,4]
[1]Department of Coastal Sciences, The University of Southern Mississippi, Ocean Springs, MS, USA
[2]Departments of Pediatrics & Department of Microbiology and Immunology, Miller School of Medicine, University of Miami, Miami, FL, USA
[3]Singing River Health System, Pascagoula, MS, USA
[4]University of South Alabama, Mobile, AL, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Autochthonous Pathogens
Allochthonous Pathogens
General Treatment Principles
Treatment Considerations for Gram-Positive Organisms
Future Directions
Bibliography

Glossary

Allochthonous Exogenous, alien or nonindigenous; arising from another source or medium.

Ambient Being of the surrounding area or environment.

Archaea One of three domains on Earth, including the *Bacteria* and *Eukarya*. *Archaea* are prokaryotes that do not have peptidoglycan cell walls; they lack membrane-bound organelles (e.g., nucleus, mitochondria, endoplasmic reticulum, chloroplasts), possess 70 S ribosomes and have ether-linked lipids in their membranes.

Autochthonous Indigenous, native, arising from within.

Bacteria One of three domains on Earth, including the *Archaea* and *Eukarya*. *Bacteria* are prokaryotes that possess peptidoglycan cell walls; they lack membrane-bound organelles (e.g., nucleus, mitochondria, endoplasmic reticulum, chloroplasts), possess 70 S ribosomes and have ester-linked lipids in their membranes.

Biodiversity The richness or complexity of life forms in an ecosystem, biome or on Earth itself.

Commensal organism An organism participating in a symbiotic relationship in which one species derives some benefit while the other is unaffected.

Ectotherm An organism that controls body temperature through external means.

Endotoxin The lipid component (lipid A) of the outer membrane lipopolysaccharide (LPS) of all gram-negative bacteria. Endotoxin is released into a host or the environment when the cell lyses and its outer membrane breaks up.

Epidemic The incidence of disease above the normal or endemic incidence.

Eukarya One of three domains on Earth, including the *Bacteria* and *Archaea*. *Eukarya* possess membrane-bound organelles (e.g., nuclei, mitochondria, chloroplasts), histones associated with their DNA and 80 S ribosomes in their cytoplasm. Plants and animals are eukaryotic.

Exotoxin Any toxin that is secreted into the cell's immediate environment. Most exotoxins are proteins, and they are made by both gram-negative and gram-positive bacteria.

Facultative organism An organism that is capable of growth both in the presence and absence of oxygen.

Foodborne disease A disease that is caused by the ingestion of pathogens conveyed by food.

Food intoxication Illness caused by the ingestion of food that contains a toxic substance.

Hemolysin A proteolytic enzyme that lyses red blood cells.

Lysogenic conversion Insertion of bacterial virus (bacteriophage) DNA into the chromosomal DNA of its bacterial host thereby conferring one or more new traits on the host.

Nosocomial Infections (and disease) that are acquired in clinical settings (e.g., hospitals, outpatient clinics, emergency rooms, physician offices).

Opportunistic pathogen Any pathogen that accidently acquires entrance to a host and then only causes disease if one or more risk factors are present in the host.

Pandemic An epidemic of world-wide proportions.

Pathogenesis The production or development of a disease, specifically the cellular reactions and other pathologic mechanisms occurring in the progression of the disease.

Pathogenic The ability of a species to cause disease. However, because pathogenesis is typically caused by one or more than one virulence factors produced by one or more genes, any given pathogenic species will often display different degrees of pathogenesis.

Pathogenicity island A cluster of virulence genes (and sometimes cryptic genes and other small genetic elements) flanked by direct repeats, insertion sequences or tRNA genes such that the clusters are easily transmitted to other bacteria via a process called *horizontal gene transfer.*

Plasmid A circular, double-stranded DNA molecule containing specialty genes that, in general, are not essential for survival of the host bacterium or genes that are cryptic (unknown). Plasmids can replicate autonomously or integrate into and replicate with the chromosome. Plasmids are smaller than the chromosome, on average 5% the size of the chromosome.

Point source A single, identifiable localized source of something.

Quorum sensing A chemical mechanism used by bacteria to measure their population density. When the chemical signals reach a certain level, special genes are expressed.

Sentinel An indicator whose presence is directly related to a particular quality in its environment at a given location.

Sequences The order of nucleotides in a specific length of DNA or RNA.

Virulence The degree of pathogenicity. Virulence is a compilation of toxins, hemolysins, proteases and lipases that may not be possessed by all strains of a pathogenic species.

Waterborne disease A disease that is transmitted by water.

Zoonosis An animal disease transmissible to humans under natural conditions or a human disease transmissible to animals.

Definition of the Subject and Its Importance

Several bacteria that cause human disease can be found in the ocean. The actual incidence of bacterial disease that results from seawater or seafood is not precisely known but is thought to be relatively low in the USA, although some diseases are on the rise. Bacterial disease from the ocean is more prevalent worldwide, especially in developing countries and in countries that derive most of their protein from seafood. Compared to the viruses, bacteria account for a much lower incidence of disease emanating from the ocean, both in the USA and worldwide. However, it is important to understand and mitigate bacterial disease from the ocean, because of such environmental pressures as global warming, antibiotic resistance, pollution, breakdowns in sanitation (e.g., Haiti after the earthquake) and tourism.

Introduction

Myriad bacteria reside in the ocean and most (>99.9%) have not been isolated and are known only by their unique molecular signatures (e.g., 16 S rRNA sequences) [35, 59]. These bacteria belong to the domain *Bacteria* which exhibits extensive *biodiversity*, only exceeded by the biodiversity within the domain *Archaea* and the viruses. The vast majority of bacteria in the ocean do not cause disease and Archaea are not known to cause any human disease [39]; however, a very small number (percentage) of bacterial species found in the ocean cause *pathogenesis* in plants and animals. Some of these pathogens are indigenous to the ocean and are defined as *autochthonous*. Others (*allochthonous* bacteria) are of exogenous or terrestrial origin and are introduced to the ocean via surface runoff, rivers and streams, atmospheric fallout and ocean disposal of wastes (intentional or accidental); the chapter on enteric viruses (See S.NO.12, J. Woods, Waterborne Diseases of the Ocean, Enteric Viruses, this volume) has a nice discussion of allochthonous sources. This chapter will focus on those allochthonous and autochthonous bacteria that cause *waterborne* and *foodborne* disease in humans. Some also cause disease in marine animals (and a few in marine plants) but nonhuman diseases will not be addressed in this chapter.

Autochthonous human pathogens have evolved in the ocean and for largely unknown reasons have the ability to infect and cause disease in humans. Almost all of these diseases result from ingesting seawater, eating seafood, or broken skin contact (swimming, wading, or working in seawater). The allochthonous human pathogens are transients in the ocean with varying abilities to survive in seawater; all of the allochthonous pathogens infect humans via the same routes as autochthonous pathogens, i.e., contaminated seawater, seafood or broken skin contact. There are a few bacteria that may fit both definitions, and there is no current scientific consensus about the place of these bacteria. Examples of these ubiquitous bacteria include the enterococci [159], *Staphylococcus aureus* (see Allochthonous Pathogens in this chapter) and *Pseudomonas aeruginosa* [114].

In the following pages, allochthonous and autochthonous marine bacteria that cause disease(s) in humans will be discussed in terms of their biology, ecology, pathogenesis, and epidemiology. Disease treatment will also be discussed but not in clinical detail.

Autochthonous Pathogens

Autochthonous pathogens generally cause one or more diseases in marine animals and if the disease(s) is transmissible to humans it is classified as a *zoonosis*. It some cases, humans can transmit disease to marine animals.

Vibrio

Vibrios, and specifically *Vibrio cholerae*, were first observed by Pacini in 1854 and later isolated in pure culture from cholera patients by Robert Koch in 1883 [8, 27]. Today, the Vibrios comprise a large genus (>80 species) and belong to the *Gammaproteobacteria* [47]. They are gram-negative rods, often slightly curved (e.g., *V. cholerae*) and most are motile in liquids by means of one or more polar flagella; on solid surfaces they are motile by means of lateral flagella. Unlike most bacteria, Vibrios possess two circular chromosomes (one large and one small) which are of relevance to this chapter because of the distribution of *virulence* genes on the two chromosomes. The large chromosome is usually referred to as Chromosome 1 (Ch1) and it tends to contain housekeeping genes (DNA replication, transcription, translation, flagellar synthesis, metabolic pathways). Chromosome 2 (Ch2) tends to contain accessory genes (pathogenicity, antibiotic resistance, host defense avoidance, survival in adverse environments). Most of the Vibrios can metabolize a large number of organic compounds, including sugars, amino acids, fatty acids, carbohydrates, proteins, lipids, alcohols, and selected aliphatic and aromatic hydrocarbons. Indeed, prior to the development of molecular biology methods, the Vibrios were largely identified and classified by these diverse and extensive metabolic traits [22].

In addition to metabolic diversity, the pathogenic Vibrios also possess an array of *exotoxins*, proteases, transport proteins, attachment mechanisms, and lipases that act as virulence factors. Since Vibrios are gram-negative, they also possess *endotoxin* which differs in toxicity from strain to strain. Indeed, the Vibrios are well equipped to cause disease in their hosts – accidental or otherwise. In general, the Vibrios are opportunistic pathogens – for both humans and marine animals – and they cause systemic infections, skin lesions and gastroenteritis. In fish, the infections often lead to hemorrhagic skin lesions (known as vibriosis) and pathology in the liver, spleen, and kidney. In humans, the diseases arise from contamination of cuts and other skin lesions with seawater (and to a lesser extent marine animals, e.g., stingray barbs and barnacles) and from ingestion of raw, undercooked or cooked but re-contaminated seafood and seawater. Depending on the species, skin infections can remain localized or become systemic and ultimately fatal. Oral ingestion of seafood and seawater leads to various degrees of gastroenteritis and, in some cases (e.g., *V. vulnificus)* life-threatening systemic infections. Although 12 *Vibrio* spp. are considered to be human pathogens, the predominant human pathogens are *V. cholerae, V. parahaemolyticus* and *V. vulnificus*. Table 1 presents a recent compilation of *Vibrio* diseases in the USA. In addition to the eight species listed in Table 1, the other human pathogens include *V. cincinnatiensis, Photobacterium damselae* (initially named *V. damsela*), *V. furnissii*, and *V. harveyi*; these four Vibrios, along with *G. hollisae* and *V. metschnikovii*, will not be discussed in this chapter, due to their low incidence of disease in humans.

All of the Vibrios require NaCl to grow properly, although this requirement is minimal for some species, e.g., *V. cholerae* and *V. mimicus*. Most if not all Vibrios use a sodium motive force to drive their polar flagella [94, 157] and at least a few species appear to use a sodium motive force to make ATP [37, 138]. Lateral flagella are produced by Vibrios when they grow on solid surfaces; these flagella are driven by a proton motive force [94] and are responsible for swarming. In addition to a requirement for sodium, most Vibrios are mesophiles and, therefore, do best in warmer waters. Recently, it has been suggested that global warming might increase the incidence of Vibrios worldwide [25, 28]. Interestingly, laboratory-confirmed infections with *Vibrio* spp. began to exceed those from *Salmonella*, Shiga toxin-producing *Escherichia coli* O157, *Campylobacter* and *Listeria* in late 2000 and this US trend has increased through 2008 [17]. In 2009, the overall rate of foodborne disease

Human Bacterial Diseases from Ocean. Table 1 Number of *Vibrio* illnesses (excluding toxigenic *V. cholerae*) in the USA, 2008[a]

Species	Patients	Deaths	Isolates	Specimen[b]
V. alginolyticus	99	1	99	Wound and other[c]
V. cholerae (nontoxic)	50	2	50	Stool and blood
V. fluvialis	29	3	29	Stool and wound
V. hollisae[d]	4	1	4	Stool and wound
V. metschnikovii	1	0	1	Blood
V. mimicus	32	0	32	Stool and other
V. parahaemolyticus	270	4	270	Stool and wound
V. vulnificus	85	24	94	Blood and wound
Species not identified	23	0	24	Stool and other
Multiple species	6	0	13	Wound and stool
Total	**599**	**35**	**616**	**Stool most frequent**

[a] Data reported to the Cholera and Other *Vibrio* Illness Surveillance (COVIS) system maintained by the CDC (see http://www.cdc.gov/nationalsurveillance/PDFs/Jackson_Vibrio_CSTE2008_FINAL.pdf)

[b] The predominant specimen types with the most predominant listed first. Additional specimens, including actual numbers, can be seen in the COVIS report[a]

[c] "Other" consists of ear, sputum urine, and other specimens

[d] Reclassified to *Grimontia hollisae* in 2003 [137]

Human Bacterial Diseases from Ocean. Table 2 Temperature and salinity preferences for the pathogenic Vibrios[a]

Vibrio species	Temperature		Salinity	
	Optimum	Range	Optimum	Range
V. cholerae	25°C [24, 26]	17–40°C [15, 31]	2–14 ppt [26, 31]	<1–60 ppt [15]
V. parahaemolyticus	38°C [27, 33]	8–45°C [16, 33]	17–23 ppt [19, 27, 44]	<1–96 ppt [33]
V. vulnificus	30°C [39]	9–40°C [21, 39, 40, 106]	5–10 ppt [25]	5–35 ppt [25, 40, 106]

[a] Bracketed numbers are supporting citations found in the *Primary Literature*

caused by Vibrios was 0.35 per 100,000 population [18]. The relative rates of the other four pathogens have either remained level (*Salmonella*) or shown a decrease (*Shigella* rates decreased 40% and STEC O157 decreased 25%) while Vibrio rates have increased by 47% [17]. Some have suggested that because this trend began just before the El Niño years 2002–2003 and continued through the El Niño years 2006–2007, it was caused by global warming. Intriguing as it may be, this observation is circumstantial and cannot be confirmed. However, in 1991 there was an outbreak of cholera in Peru, a country that had not seen this disease for over 100 years [136]. The source of the 1991 outbreak was never determined and two more outbreaks occurred in 1993–1995 and 1997–1998. Gil et al. [56] very nicely demonstrated that the Peru outbreaks in summer 1998 correlated (linear regression, $P < 0.001$) with the sea surface temperature peak associated with the strong El Niño that year. Colwell et al. [25] and [28] have made similar observations for the Bay of Bengal. In 2004, an outbreak involving cultured raw oysters and *V. parahaemolyticus* occurred aboard a cruise ship in Prince Edward Sound [96]. Mean water temperatures had been increasing in Prince William Sound (0.21°C per year) since 1997, and in 2004, for the first time, mean daily temperatures in the sound did not drop below 15°C [96]. Johnson et al. [74] reported that when temperatures drop below 15°C *V. parahaemolyticus* and *V. vulnificus* are no longer culturable from water and sediment in Mississippi Sound; similar observations have been reported by others, suggesting that 15°C is a limiting temperature for many of the Vibrios. Clearly, it is beginning to look like increasing ocean temperatures are increasing the incidence of Vibrio disease and as more data become available this Vibrio incidence-climate link may be substantiated. Temperature and sodium preferences for the three principal pathogenic Vibrios are shown in Table 2. Finally, in the past decade, scientists are using remotely-sensed satellite data, including temperature and salinity, to predict human health risks from pathogenic vibrios in water and oysters [90, 117].

Vibrio alginolyticus This *Vibrio* is very common in estuarine and marine waters worldwide; in fact, it is unusual not to isolate *V. alginolyticus* when culturing coastal and estuarine water samples. It is known for its ability to swarm on isolation media [141] and it often overgrows desired isolates. Recently, it was demonstrated that *V. alginolyticus* is capable of quorum sensing and that QS is responsible for biofilm formation (swarming is a prelude to this phenomenon) and development of virulence factors involved with fish disease [64].

The most common disease caused by *V. alginolyticus* in humans is wound infection. Of the 99 infections reported to the CDC in 2008 (Table 1), 64 isolates came from wound infections, 5 from blood, 3 from stools and 27 from other specimens. Most infections are mild and self-limiting, although *V. alginolyticus* has been demonstrated in a few cases of severe necrotizing fasciitis that involved patients at risk because of cirrhosis [66]. Cases of gastroenteritis caused by *V. alginolyticus* are rare [140], as indicated in Table 1. This Vibrio has also caused eye infections [86].

Vibrio cholerae During the early cholera pandemics of the 1800s, *V. cholerae* was responsible for millions of deaths and was feared as the most dangerous waterborne human pathogen known [27, 76]. It still causes several thousand deaths annually [153] and a small number of cases occur in the USA each year (see COVIS report cited in Table 1). The World Health Organization recently reported that in 2009 the number of cases worldwide (45 countries reporting) was 221,226 with 4,946 deaths (2.24% case fatality rate) [153]. The number of cases in the USA in 2008 was nine and none of these patients died (see COVIS report, Table 1). Although molecular detection is very effective and widely used [102] and the genomes of several strains have been fully sequenced, for documentation and epidemiology purposes isolates are still serogrouped by means of their O antigens. The predominant serogroups causing human disease worldwide are O1 and O139; in the USA, the CDC tests for serogroups O1, O75, O139, and O141 (COVIS report cited in Table 1).

The classic disease caused by *V. cholerae* is a rapidly developing, profuse, watery diarrhea that is usually accompanied by severe dehydration. The cholera diarrhea is often called a "rice-water" stool because as the disease progresses the frequent stools are little more than water containing flecks of mucous abraded from the intestinal mucosa. When death occurs, it is because of the severe dehydration (water loss) and acidic coma (loss of sodium bicarbonate to the diarrhea). Historically, the death rate often exceeded 50% and death could result in as few as 24–36 h after onset of symptoms. As noted above [153] the death rate for cholera is now below 5% and most cases of cholera today are little more than a transient diarrhea. In addition to classic cholera, *V. cholerae* can also cause self-limiting gastroenteritis and wound infections; it has also been implicated in fish disease (eels and ayu). An interesting and illustrative human case occurred after Hurricanes Katrina and Rita and involved a Louisiana couple who consumed properly cooked (boiled) shrimp that had been placed on the ice used by them to transport the raw shrimp. The husband developed a severe case of cholera that caused renal, pulmonary and cardiac failure. He was given ciprofloxacin and aggressive rehydration therapy and he did not die. Of interest is that the husband had a history of common Vibrio risk factors – history of high blood pressure, alcoholism, diabetes, brain tumor, and renal failure that required frequent dialysis. His wife had mild diarrhea and was treated (ciprofloxacin and rehydration) as an outpatient. *V. cholerae* O1, serotype Inaba, biotype El Tor was isolated from both patients [15].

The main virulence factor associated with cholera is an exotoxin known as the cholera toxin (CTX). CTX is not produced by all strains of *V. cholerae*, and it is a protein that is composed of two subunits – one A subunit and five B subunits. The *ctx* genes that encode for these two subunits, *ctx*A and *ctx*B, are actually genes in a virus (a bacteriophage or phage) that infects *V. cholerae* and establishes a lysogenic relationship with its host [148]. The CTX phage that infects *V. cholerae* is a lambdaphage; and instead of producing more phage when it infects its host, this phage inserts its DNA into the chromosome (Ch1) of the host which then allows the host to produce CTX – a process called *lysogenic conversion*.

V. cholerae produces other virulence factors, including the toxin-coregulated pilus (TCP) produced by the *tcp*A gene, repeat-in-toxins (RTXs), and a heat-stable enterotoxin (NAG-ST) that is related to the heat-stable enterotoxin produced by *E. coli* [135]. The TCP is thought to be necessary for intestinal colonization by *V. cholerae* and may also serve as an attachment site for CTX phage [148]. Type 2 secretion systems are present in *V. cholerae*, and these systems provide a physical conduit for the bacteria to secrete toxins into their host cells (see secretion system discussion in "*Vibrio parahaemolyticus*" section).

In closing, prevention and treatment of cholera deserves mention. Cholera has a very low incidence in developed countries, primarily because of good sanitation, but this is not the case in developing countries [68]. In both developed and developing countries, cholera is usually a self-limiting disease requiring minimal treatment other than rehydration and electrolyte replacement; antibiotics are not usually administered. In developing countries where safe drinking water is not readily available, filtration of drinking water is very effective in preventing cholera. A very simple but highly effective filtration procedure was pioneered in Bangladesh [26]. The procedure involves using folded sari cloth (eight layers of old sari) to filter the drinking water; this procedure removes zooplankton to which the cholera bacilli are attached [26].

Cholera in Haiti

In January, 2010 Haiti was ravaged by a major earthquake that had an epicenter approximately 25 km from the capital city of Port-au-Prince. On October 21, 2010 cholera was confirmed in Haiti by the CDC; the causative agent has been identified as *V. cholerae* O1 serotype Ogawa. As of October 14, 2011, the Hatian Public Health Ministry reported that over 473,649 people have been infected and 6,631 patients have died. The disease peaked in early 2011 but cases and deaths continue. Interestingly, Haiti had not reported a cholera case for over a century and the source of the outbreak was, early on, not known. Clearly, the lack of hygiene, safe drinking water and safe food that followed the earthquake and continues to exist has contributed to the onset and continuation of cholera in Haiti. However, recent molecular evidence reported in the CDC journal "Emerging Infectious Diseases" demonstrated that a UN peacekeeping force from Nepal introduced the cholera strain into the Meille River.

Vibrio fluvialis The incidence of disease caused by *V. fluvialis* is thought to be very low, both in the USA (Table 1) and worldwide. This *Vibrio* primarily causes enteric disease but can cause wound infections and, rarely, other extraintestinal infections.

Vibrio mimicus This *Vibrio* is similar to *V. cholerae* in many ways and early on was thought to be an atypical *V. cholerae* – hence the species name "mimicus." The primary disease caused by *V. mimicus* is gastroenteritis and some strains carry *ctx* genes as well as other virulence factors found in *V. cholerae* and other Vibrios.

Vibrio parahaemolyticus First isolated in Japan from a gastroenteritis outbreak traced to the consumption of shirasu, a popular fish product (272 cases and 20 deaths; [54]), *V. parahaemolyticus* is the most common cause of foodborne disease in countries that consume high quantities of seafood [61]; in the USA, it is the most common cause of seafood-borne disease (see Table 1). As is the case with all the Vibrios that cause human disease, *V. parahaemolyticus* causes gastroenteritis and wound infections, and it is the most common cause of Vibrio disease in the USA (Table 1). *V. parahaemolyticus* can be isolated from most forms of seafood but is most commonly associated with shellfish [12].

The first *V. parahaemolyticus* pandemic began in 1996, and it continues to involve three major serotypes – O3:K6, O4:K68, and O1:K untypable [23]. Recent outbreaks are described in Table 3.

All *V. parahaemolyticus* strains possess the *tlh* (thermolabile hemolysin) gene, and this gene is frequently used to rapidly detect and confirm identification of the species. However, *tlh* is not unique to *V. parahaemolyticus*; *V. harveyi*, *V. alginolyticus*, and *V. fischeri* have also been shown to contain a *tlh* homologue or a related hemolysin gene [72, 134, 149].

When first discovered, it was shown that *V. parahaemolyticus* had the ability to hemolyze red

Human Bacterial Diseases from Ocean. Table 3 Recent outbreaks and cases of *V. parahaemolyticus* gastroenteritis

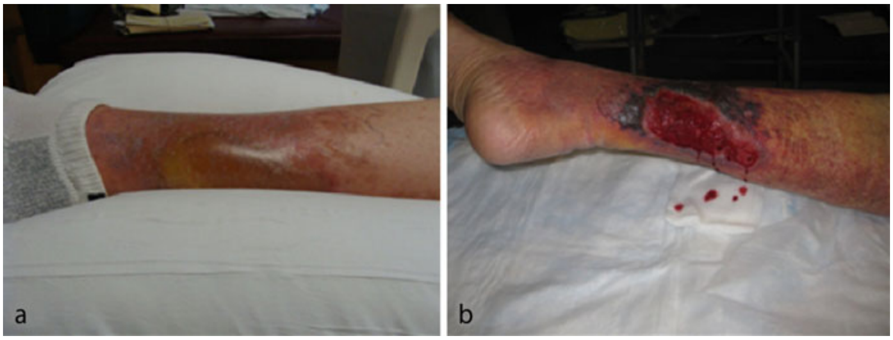
Date	Location (ref.)	Source	Cases/deaths	Serotype	El Niño
1997	Pacific North West [12]	Oysters	209/1	O1, O4, O5	Yes
1998	Gulf of Mexico, NE, Pacific NW [32]	Oysters	416/?	O3:K6	Yes
1998	Chile [62]	Shellfish	~300/?	O3:K6	Yes
1998	Japan [61]	Seafood	12,318/?	O3:K6	Yes
2004 and 2005	Chile [62]	Seafood	~5,100/?	O3:K6	No
2004	Alaska [96]	Oysters	62/0	O6:K18	No
2006	New York, Oregon, Washington [16]	Oysters	177/0	O4:K12	Yes
2006	Chile [62]	Shellfish	900/?	O3:K59	Yes

blood cells in a special culture medium called Watsumaga agar; it was later shown that this hemolysis, called the Kanagawa reaction, was mediated by a hemolysin called thermostable direct hemolysin TDH [67]. Most pathogenic *V. parahaemolyticus* strains possess the *tdh* gene and/or the *trh* (thermostable related hemolysin) gene [73], although some pathogenic strains contain neither *tdh* nor *trh* [98]. The *tdh* genes are located in a *pathogenicity island* on Ch2 [91]. In general, environmental and food isolates do not contain tdh and trh ([73] and [74]).

Gram-negative bacteria possess a fascinating injection apparatus called secretion systems. These allow bacteria to inject various substances into the cells of *Eukarya***.** As is the case for many bacterial virulence factors, especially those associated autochthonous pathogens, such as the Vibrios; microbiologists understand how these factors function in human pathogenesis but do not know the function of these factors in the environment. To date, six secretion systems (T1SS, T2SS, T3SS, T4SS, T5SS and T6SS) have been described. In 2003, the sequenced genome of *V. parahaemolyticus* RIMD2210633 was shown to contain two different T3SS genes – T3SS1 and T3SS2 [91]. T3SS1 is located in a pathogenicity island on Ch1, T3SS2 is in a Ch2 pathogenicity island, and it is now known that there are two different T3SS2 – T3SS2α and T3SS2β [103].

Makino et al. [91] identified at least 50 other genes in *V. parahaemolyticus* that may be involved with pathogenesis. Some of these additional virulence factors include urease [112], attachment mechanisms [60, 130], ToxR [30], and RTX toxin [91].

Vibrio vulnificus One of the more recently discovered pathogenic Vibrios [46, 123], *V. vulnificus*, causes serious diseases in humans and is thought by some Vibrio biologists to be the most virulent bacterium now known (death rate >50% in patients at risk [104]; the overall case fatality rate reported by CDC [97] for 1992–1997 was 39%). *V. vulnificus* is thought to cause three human diseases: primary septicemia (caused by the ingestion of raw or undercooked shellfish – especially oysters), gastroenteritis (caused by the ingestion of raw or undercooked shellfish), and wound infections (caused by contact with water, barnacles, fish barbs, and other marine objects). The literature documenting simple gastroenteritis is scant and, accordingly, some experts question this disease having a *V. vulnificus* etiology (J.D. Oliver, personal communication). Primary septicemia is a rapidly developing disease that can result in death in less than 48 h after consuming seafood containing the bacteria. The bacteria move from the intestinal tract into the blood stream and, from there, set up serious infections in tissues, especially the extremities. Wound infections also most frequently involve extremities and will cause bullas (Fig. 1). When the infection progresses into deep tissue and begins to cause death of that tissue, the disease is called necrotizing fasciitis. The infection shown in Fig. 1 did not develop into necrotizing fasciitis but is a good example of what often happens when patients



Human Bacterial Diseases from Ocean. Figure 1
Wound infection of unknown etiology, caused by a minor scratch from a barnacle. Within 36 h the scratch had developed into a bula (**a**). The bula was debrided (**b**) and did not progress into necrotizing fasciitis (Grimes DJ, Ekenna O unpublished data)

are placed on aggressive antibiotic therapy to prevent progression of the infection. In this case, the patient was given three different antibiotics (clindamycin, levafloxacin, and doxycycline) prior to culture of the lesion, and these antibiotics probably caused the etiological agent to become nonculturable, a phenomenon observed and discussed by others [55, 126].

In general, *V. vulnificus* infections in healthy individuals are often not serious, but this is not always the case. Serious and fatal infections can occur in all patients and especially in those at risk; risk factors include: preexisting liver dysfunction or disease, diabetes, alcoholism, poor circulation, and immunosuppressive drug therapy. It is imperative that such individuals at risk are properly counseled about contact with seafood and seawater, so that they do not become infected with any *Vibrio* capable of causing disease in humans.

The virulence factors associated with *V. vulnificus* are not well understood. Known factors include *V. vulnificus* hemolysin (VVH) and RT [104]. There are two genotypes of *V. vulnificus* – clinical (C) and environmental (E) – and only C causes disease in humans. All strains of *V. vulnificus* have powerful iron sequestration ability (siderophores) which allows them to out-compete other species (including humans and fish) for essential iron. Capsular polysaccharide (CPS) formation is important and only encapsulated strains are virulent [104]. In addition, endotoxin (the lipoidal moiety of LPS) is very important and may be the most important cause of shock and death from *V. vulnificus* infections [104]. The complete genomic analysis of *V. vulnificus* YJ016 revealed the presence of RTX genes and they along with siderophore genes are located on Ch2 [21]. Type IV pilins, used as adherence mechanisms by many *Gammaproteobacteria* including *V. cholerae*, are a consistent feature of *V. vulnificus* and function in biofilm formation, attachment to epithelial cells, and possibly in the colonization of oysters [109, 110]. Other putative virulence factors have been described but their role in pathogenesis is unclear.

Aeromonas

Members of the genus *Aeromonas* are primarily freshwater bacteria that cause disease in both humans and aquatic animals. The most common isolates from human clinical specimens are *A. hydrophila, A. caviae, and A. sobria*, with *A. hydrophila* being the most common. Although *Aeromonas* spp. are frequently isolated from estuaries and the coastal ocean, they may not be truly autochthonous, and human disease from *Aeromonas* does not normally result from seawater or seafood. In fact, there is controversy about the ability of *Aeromonas*, a well-known pathogen of fish and amphibians, to cause disease in humans [48]. If one accepts the literature in support of human disease, it is believed that most human disease occurs in the form of gastroenteritis resulting from the ingestion of fish, shellfish, red meats, and contaminated water [48]. *Aeromonas* gastroenteritis presents either as a cholera-like disease with watery stools or a dysenteric-like illness that can include bloody stools [48]. Wound infections can also result from contact with contaminated water, either freshwater, seawater or brackish water [14, 75, 144].

Three recent reports of *Aeromonas* infections are illustrative of this seafood-borne and waterborne pathogen. In the aftermath of the 2004 tsunami that devastated eight countries and caused an estimated 225,000 deaths, the Thai Ministry of Public Health quickly began meeting health care needs in Thailand. From December 26, 2004 (the day of impact) to January 11, 2005, 1,237 cases of acute diarrhea, 356 wound infections, 177 febrile illnesses, and 156 respiratory illnesses were reported to the MOPH; only two deaths (both from aspiration pneumonia) resulted [14]. The most common isolates from the wound infections were *Proteus* spp., *Klebsiella* spp., *Pseudomonas* spp., *Enterobacter* spp. *E. coli*, and *A. hydrophila* (two isolates); surprisingly, no Vibrios were isolated, as was the case with Hurricanes Katrina and Rita.

Eighty-two strains of presumptive *Aeromonas* spp. were isolated from 250 frozen freshwater fish (Tilapia, *Oreochromis niloticus niloticus*) intended for human consumption and purchased in local markets in Mexico City. The isolates were identified with standard molecular techniques (16 S rRNA) and 88.3% were placed in two species – *A. salmonicida* (67.5%) and *A. bestiarum* (20.9%). The remaining isolates were identified as *A. veronii* (5.2%), *A. encheleia* (3.9%), and *A. hydrophila* (2.6%). The authors noted that this was one of the first major *Aeromonas* studies conducted in Mexico, and further noted that their results demonstrated the need for concern over putative pathogens with antimicrobial resistance and known virulence factors being present in food meant for human consumption [9].

Finally, a study of the prevalence of *A. hydrophila* in marketed seafood (fish and prawns) was conducted in land-locked city in South India by Vivekanandhana et al. [145]. Random samples of seafood (536 fish and 278 prawns) were collected from several vendors in a popular seafood market, and fish showing visible spoilage, injury or disease were avoided. Overall, 180 fish samples (33.6%) and 49 prawn samples (17.6%) contained *A. hydrophila*. The authors attributed the incidence to temperature abuse, fly contamination from a nearby sewage treatment plant, and the ability of *A. hydrophila* to grow at refrigerator temperatures (4–7°C); and they further noted that *A. hydrophila* is a pathogen of emerging importance [145].

Aeromonas Soft-Tissue Wound Infection

In late summer of 2010, a 7-year-old boy suffered a large and complex laceration injury to the right calf while recreational boating on a coastal river. He was brought within 2 h to the hospital, after initial first aid in the field. The wound was thoroughly washed and cleansed at surgery and closed with over 30 fine stitches. He was discharged the next day in stable condition and on oral antibiotics.

He failed to take the prescribed antibiotics because of nausea and vomiting (no diarrhea or fever). On readmission 6 days later, the wound was infected. Wound culture showed a rapid growth of a gram-negative rod (GNR) that was beta-hemolytic on blood agar, and also grew on MacConkey and chocolate agars. It was oxidase positive, mucoid, catalase positive, and motile.

The organism was confirmed to be *Aeromonas hydrophila*. It was found to be sensitive to second and third-generation cephalosoprins, quinolones, tetracycline, trimethoprim/sulfamethoxazole, and aminoglycosides. He responded well to intravenous ceftriaxone (a third-generation cephalosporin), local wound care, and later to applied skin graft to the injured calf. He made a full recovery.

The infection was most likely caused by wound contamination from the freshwater (river), and occurred because of a combination of factors: premature surgical closure of wound, and inability of the patient to take prescribed oral antibiotics to which the organism was sensitive.

Edwardsiella

There are three species of *Edwardsiella*, and the one that causes human disease is the opportunistic pathogen *E. tarda*. *E. tarda* causes a Salmonellosis-type enteritis in humans and typically derives from freshwater and freshwater animals (e.g., pet turtles). Infections from marine sources are unknown. *E. ictaluria* is a serious fish pathogen, often associated with septicemia in catfish (especially in aquaculture of catfish), but it is not known to cause disease in humans or marine fishes.

Yersinia

Several *Yersinia* spp. cause diseases in humans (Bubonic plague or Black Death, pseudotuberculosis, enteritis, extraintestinal complications) and in fish, including marine fish (salmonids). However, none of the marine fish pathogens are known to cause human disease and will therefore not be discussed here.

Brucella

Long known to cause disease in terrestrial animals and humans, these zoonotic bacteria are now known to exist in the ocean. The classic *Brucella* spp., *B. abortus*, *B. melitensis* and *B. suis*, cause brucellosis or undulant fever in humans and domestic animals. In domestic animals, the disease outcome is often abortion because the bacteria prefer to metabolize mesoerythritol which is found in the uterus and fetus of animals but not in humans. In humans, symptoms are general and include fever, chills, malaise, with heavy sweating, and high fever at night.

The three classic *Brucella* species have not been reported in marine mammals. Instead, it appears that marine mammals carry their own *Brucella* spp. [71], and they have caused three naturally acquired cases of human disease. Two cases were reported in Peru in 2003 (Sohn et al. 2003) and one in New Zealand [95]. The three *Brucella* human isolates were characterized with several molecular methods, and all three were found to share a common genotype with previously reported marine mammal *Brucella* spp. [152]. Representing 173 animals and one human patient were analyzed using a molecular method called multiple loci VNTR (variable number of tandem repeats), and the authors targeted 16 genetic loci (MLVA-16) that had been shown to be highly descriptive for *Brucella* spp. [93]. The study included two new species isolated from marine mammals, *B. ceti* and *B. pinnipedialis*, and concluded that these two species cluster into three distinct clades. Interestingly, the three isolates described by Whatmore et al. [152] did not cluster within the three clades but were, however, closely linked to the three marine mammal groups [93].

Enterococcus

Members of the genus *Enterococcus* are largely commensal colonizing organisms of the gastrointestinal tract of humans and warm-blooded animals and are commonly recovered in their feces [49]. These organisms are gram-positive facultative anaerobes that do not form spores but are capable of survival and growth in a wide variety of environmental conditions. These include tolerance of temperatures ranging from 10°C to 45°C, pH from 4.5 to 9.0 and high sodium chloride concentrations [63]. Although *Enterococcus* species have been found in many different marine and freshwater environments [85, 92, 124, 143, 158] as well as being associated with processed and fresh fish and seafood [31, 70, 99, 116, 154], these organisms are not usual pathogens for fish or marine mammals. *Enterococci* are known to be introduced into these environments by sewage contamination from known point sources, such as sewage treatment plants, and are used as indicator organisms for the probable presence of disease-producing pathogens in marine waters [147]. It is unlikely that point sources are the sole contributor of these organisms to an aquatic environment. Domestic and wild animals, water runoff from storms or agricultural sources, wind-driven sediment resuspension events, and humans utilizing the waters have all been shown to contribute to the presence of *Enterococcus* species in aquatic environments. ([1, 33, 42, 43, 107]; Rebarchik DM, Grimes DJ unpublished data).

The principal pathogenic *Enterococcus* in humans, *Enterococcus feacalis* and *E faecium*, are among the *Enterococcus* species isolated from aquatic environments; however, *Enterococci* found in marine settings have not been linked directly to the onset of human enterococcal infections. In general, these organisms are primarily associated with serious, often fatal, nosocomial infections, including postsurgical wound infections, endocarditis, urinary tract infections, and sepsis; and they are currently emerging common pathogens [49]. Enterococci lack significant virulence factors associated with disease but are intrinsically resistant to many antibiotics currently in use. These bacteria, especially *E. fecium*, are known to easily acquire antibiotic resistance genes from other microorganisms encountered in their environment. In addition to its importance as an indicator organism, the significance of *Enterococcus* in a marine-water setting is the increased likelihood that the organisms will be exposed to other microorganisms from which they might acquire antibiotic resistance genes, thus adding to the difficulty of treating an already challenging infection.

Streptococcus

Numerous species of fish are susceptible to infection by members of the genus *Streptococcus*. Although these infections are not common, when they do occur, it is often in an aquaculture setting and can be responsible for significant mortality and large economic costs. One species that is responsible for such infections, *Streptococcus iniae*, is a primary pathogen for fish that can also infect humans. This organism was first isolated from infected freshwater dolphins in 1996 [113] and has subsequently been associated with sporadic infections in multiple fish species [41, 77, 80, 115]. It was recognized as a human pathogen in the mid-1990s with several documented infections in North America [11, 150], and later in Japan [84]. *S. iniae* infections in humans present as fever and cellulitis, often with bacterimia, and can be treated with intravenous penicillin and gentamicin [84].

S. agalactiae (Lancefield group B) is a significant human pathogen especially in newborn infants where it can cause sepsis, pneumonia, and meningitis; and in pregnant women where it is associated with urinary tract infections. This organism has been linked to disease outbreaks and some massive kills in several fish species [121]. Investigations performed to type the bacteria isolated from infected fish and environmental samples indicated that sewage contamination was a likely source for the infections in fish [121, 125]. There is no evidence linking fish or a marine environment with human disease.

Listeria

The genus *Listeria* consists of six species including two that are recognized as pathogens. *L. monocytogenes* is an important human pathogen, and *L. ivanovii* is an animal pathogen that may very rarely infect humans but is not associated with marine related infections. *L. monocytogenes* are short, motile, gram-positive rods that may appear as coccobaccili. They are hardy organisms that grow as facultative anaerobes in a wide range of temperatures (from 1°C to 45°C), pH (pH 4.3 to pH 9), and they tolerate high salt concentrations. They are ubiquitous organisms that are commonly found in soil and water, on vegetation, and decaying matter and excreted in feces of humans and animals [89]. *L. monocytogenes* is the causative agent of listeriosis, a serious but rare infection caused by eating food contaminated with the bacteria. Multiple types of food have been associated with *Listeria* infections, classically soft cheeses made from unpasteurized milk, meat and processed meat products, and fish and shell fish [87]. Food may become contaminated before, during, or after preparation; and the usual measures for prevention of growth of contaminating organisms, low temperatures, extremes of pH, and high salt, are ineffective against *Listeria*. Included among the marine sources implicated in food contamination are crab meats and dips, lobster and shell fish, and many varieties of fish, especially those smoked or processed [6, 36, 81–83]. In a recent report from the CDC addressing the incidence of infection with pathogens that are commonly transmitted via a food borne route, of the greater than that 17,000 culture-proven infections, Listeria accounted for less than 1% of cases with an incidence of 0.34 per 100,000 [18].

Listeriosis occurs primarily in pregnant women, newborn infants, elderly patients, and patients who are immunocompromised, and in all but the newborns; the infections result in an initial mild influenza-like illness that may progress to sepsis and meningitis. In pregnant women, there is an increased risk of miscarriage, and in newborns the infection is associated with a high mortality, up to 40%, and long-term side effects [7, 120]. Adults and children acquire the disease after ingestion of contaminated foods, whereas newborns acquire infecting organisms either transplacentally or via an ascending route during labor and delivery. Onset of the disease varies with population and exposure routes, and it may be from days to weeks. The mortality rates for infections in these populations are high, up to 25%. *Listeria* are facultative intracellular pathogens, a characteristic that contributes to the severity of listeriosis. After being phagocytosed, the bacterium utilizes unique virulence factors to spread from cell to cell without an extracellular state, thus evading the humoral immune response. Included among these factors are internalins (inlA, inlB, inlC) that facilitate attachment to the cells. Once inside the cell, listerolysin O and phospholipase C enzymes act to release the bacteria to the cytosol where the bacteria utilizes a protein, ActA, to coordinate the assembly of actin into a "tail" that propels the bacterium across one cell and into the next. *Listeria* infections can be treated with common antibiotics, such as ampicillin, ciprofloxacin, linezolid and azithromicin.

Mycobacterium

The genus *Mycobacterium* consists of numerous aerobic, nonmotile, non-spore-forming, acid-fast rods that occur widely in nature. These bacteria range from soil-dwelling saprophytes to pathogens of humans and animals. Of the greater than 20 human pathogenic *Mycobacterium* species, few have been associated with infections from or in a marine environment [155]. *Mycobacterium marinum* (synonym *M balnei*) is the primary aquatic pathogen and is a free living organism found throughout the world in both fresh and marine water environments. It was first discovered as a pathogen for fish, causing skin nodules and ulcers, in the mid-1920s and has since been recognized as a natural pathogen for other *ectotherms*, such as frogs. Since the early 1950s [88] it has been recognized as a cause of human infections first described as causing superficial skin lesions, nodules referred to as "swimming pool granulomas," in children who swam in contaminated pools. Like most other human pathogenic Mycobacteria, *M marinum* grows slowly, having a typical incubation period of 2–4 weeks before cutaneous lesions appear, but occasionally may progress rapidly. Unlike other human pathogenic Mycobacteria, it optimally grows at lower temperatures, 30–33°C, and is usually inhibited at 37°C, helping to explain the typical location of the lesions on the extremities and the usual lack of systemic involvement.

Currently *M marinum* is the most common cause of atypical *Mycobacterium* infection in humans with a reported incidence of 0.27 cases per 100,000 adults [128] Infection with these organisms can occur at any age, but it usually occurs in adults who handle fresh- or saltwater fish or fish tanks, and is now rarely associated with swimming pools, as proper construction and chlorination has eliminated this source [40]. Human exposure now primarily comes through inoculation of existing skin abrasions while handling fish or their tanks, or directly through fish bites or puncture wounds from fins. Local trauma is an important factor in establishing *M marinum* infections and their sequelea. Infections obtained after inoculation of an existing abrasion or a direct puncture manifests as a localized nodule or granuloma at the site of bacterial entry that may develop into an ulcer or progress to involve nearby lymph nodes, sporotrichotic lymphangitis. In healthy individuals it rarely progresses to involve bones, joints, or other systemic sites. Immunocompromised states increase the risk for becoming infected and can be associated with more aggressive systemic disease [111, 151]. Diagnosis and treatment are often delayed because of a lack of suspicion for mycobacterial involvement versus more common bacterial pathogens and are contributed to by the long incubation period.

Treatment for *M marinum* is driven by the severity of the infection [122] and ranges from oral monotherapy with minocycline, clarithromycin, doxycycline, ciprofloxacin, and trimethoprim-sulfamethoxazole for superficial cutaneous infections with susceptible organisms to combination therapies for drug resistant strains. Severe infections, including those with a sporotrichoid distribution pattern, generally require combination therapy with rifampicin and ethambutol. Surgical debridement is not usually recommended, however, other alternative topical therapies, such as cryotherapy, X-ray therapy, electrodesiccation, photodynamic therapy, and local hyperthermic therapy can be effective.

Allochthonous Pathogens

Escherichia

The first species of this genus, *Bacterium coli*, was first isolated in the late 1800s and it was proposed as an indicator of fecal contamination of water. The isolate was renamed *Escherichia coli* in 1919 and today remains in use as the fecal indicator recommended by the USEPA for freshwater [142]. Unfortunately, *E.coli* and the enterococci do not correlate with indigenous pathogens such as the vibrios [29] and enteric viruses (see s.no. 12, J. Woods, Waterborne Disease of the Ocean, Enteric Viruses, this volume). In addition to its worldwide use as a fecal indicator, *E. coli* is the most common cause of urinary tract disease in humans (ca. 90% of human UTIs): certain strains cause gastroenteritis of various degrees of severity (e.g., STEC or the Shiga toxin-producing *E. coli* O157 usually derived from meats and produce) and nonpathogenic strains such as *E. coli* K12 that laid the early foundation for much of what is known about metabolism and enzymology.

Despite extensive documentation of *E. coli* causing human disease from the consumption of contaminated raw and undercooked foods, oral–fecal transmission in public facilities such as nurseries and day care centers and contaminated drinking water, transmission from seafood and seawater is rare and does not occur in the USA with great frequency. The overall foodborne STEC O157 incidence in the USA for 2009 was 0.99 per 100,000 population, and the STEC non-O157 incidence was 0.57 [18] for many years. This disease agent will not be further discussed in this chapter, and readers interested in this important pathogen are encouraged to peruse other literature.

Shigella

Shigella spp. cause a significant incidence of foodborne disease worldwide, but they are not often acquired from the ocean. Some strains produce the powerful Shiga toxin which, like CTX, is a lysogenic conversion product. In the USA, shigellosis (bacillary dysentery) has a fairly high incidence (3.99 per 100,000) as a foodborne agent of disease [18].

Salmonella

The Salmonellae are important pathogens but are much like *E. coli* and *P. aeruginosa*, with regard to their importance as marine pathogens. There are two species of *Salmonella, S. enterica* and S. *bongori*, and these two species are comprised of many serovars and subspecies. Both *Salmonella enterica* serotype Typhi (formerly *S. typhi*) and the other gastroenteritis-causing *Salmonella* spp. are human pathogens that historically were frequently acquired from the ingestion of contaminated seafood (especially filter-feeding bivalves, such as raw or undercooked oysters) and seawater. With the advent of fecal indicator monitoring of seafood and seawater, refrigeration of seafood, sanitary surveys (especially surveys of molluscan shellfish beds), and sewage treatment in developed nations, the origin of these diseases from the ocean declined significantly. Today, diseases caused by Salmonellae are still frequent worldwide, most commonly caused by *S. enterica* subspp. *enterica*, and they are almost always foodborne in both developed and developing countries (see http://www.who.int/mediacentre/factsheets/fs139/en/; [97]). In the USA, CDC reported a Salmonella foodborne disease incidence of 15.9 per 100,000 population; this is the highest incidence of any foodborne disease, but there was no breakdown on food type, i.e., seafood incidence was not given [18].

Non-typhoid salmonellosis in humans is most commonly gastroenteritis although complications, such as septicemia, can occur; deaths are rare if the patients are kept hydrated and placed on appropriate antibiotic treatment. Virulence is largely determined by pathogenicity islands, and the non-typhoid Salmonellae contain at least 12 of these genetically mobile elements [65]. Human salmonellosis is usually acquired from food, although contaminated water can also serve as a vehicle for transmission. The Food and Agriculture Organization (FAO) of the United Nations recently published a report on the control of *Salmonella* in sustainable aquaculture, and it contains a very nice review of occurrence and survival in the aquatic environment [44].

The incidence of salmonellosis deriving from seafood in the USA is low but is probably far underreported. DePaola et al. [34] recently provided evidence that, even though the reported incidence is low, the potential for acquisition in the USA certainly exists. They conducted a 2-year study of market oysters collected twice each month from retail establishments (restaurants and raw bars, seafood markets, wholesale dealers) in nine states. In all, FDA collected 397 samples representing 258 establishments. Salmonella was detected in 8.6% of the market oysters, a rate only exceeded by *V. parahaemolyticus* and *V. vulnificus* [34].

In the FAO report [44], it was noted that many studies have shown *Salmonella* serotype Sneftenberg to be the major serotype in marine environments and raw seafood worldwide. The report further noted that *Salmonella* spp. have been isolated from many marine mammals.

Morganella

Human disease from *Morganella morganii* is common (postoperative and other nosocomial) but these infections rarely emanate from the ocean. However, *M. morganii is* often associated with the decomposition of seafood; and if such seafood is consumed, scombroid fish poisoning can result. Scombroid results from histamine build up (and possibly buildup of other vasoactive amines) in the seafood as a result of histidine decarboxylation during the seafood spoilage process. This disease is a true food poisoning or intoxication, as opposed to a food infection, e.g., Salmonellosis and *Vibrio* gastroenteritis. Scombroid is probably caused by several enteric bacteria, including *Proteus*, spp. *Klebsiella pneumoniae, Hafnia alvei, Enterobacter* spp., *Serratia* spp., and *Citrobacter freundii;* in addition, *V. alginolyticus, Aeromonas* spp., and *Photobacterium* spp. are also histamine formers; and all have been isolated from spoiled fish [139]. The incidence of scombroid in the USA is thought to be common (http://www.fda.gov/Food/FoodSafety/FoodborneIllness/FoodborneIllnessFoodbornePathogensNaturalToxins/BadBugBook/ucm070823.htm), and worldwide incidence is also common. However, because scombroid is not a reportable disease, documented cases are very low, e.g., only 103 incidents involving 827 people were reported from 1968 to 1980 (see above FDA URL). The most common fish involved with this intoxication are tuna, bonito, mackerel, and *mahi mahi*, and once the amines are formed in the meat, neither cooking, canning, or freezing lowers toxicity. Onset of this intoxication is rapid (often 30 min) and symptoms include a tingling or burning sensation in the mouth, rash on the upper body, and a drop in blood pressure; nausea, vomiting, and diarrhea may also present, and hospitalization may be required.

Pseudomonas

The most common human pathogen in this genus is *Pseudomonas aeruginosa*, and, like the Vibrios, this species is metabolically very diverse. *P. aeruginosa* is a very common cause of death in third-degree burn patients, it is a common nosocomial agent of disease, is frequently resistant to most clinically useful antibiotics, it is a cause of urinary tract infections, and it is a common (and often lethal) complication of cystic fibrosis. There have been numerous reports of human *P. aeruginosa* infections occurring from various types of freshwater contact (e.g., swimming pools, whirlpools, hot tubs, atomizers), but this literature is far too extensive to be summarized here. Although *P. aeruginosa* is frequently isolated from the coastal ocean [58], the authors are not aware of any literature documenting that disease caused by *P. aeruginosa* came from the ocean. It is primarily a freshwater bacterium [114].

Campylobacter

Campylobacter spp. are frequently isolated from healthy cattle, chickens, and birds [48], and they are also associated with several foods (unpasteurized milk, poultry, shellfish, fruits, and vegetables), freshwater ponds, and streams contaminated with fecal material [69]; Campylobacters are not normally isolated from seawater. While a few reports of seawater isolations exist [3, 58], most marine isolates come from shellfish [69]. The association with shellfish is similar to that of *Salmonella*, in that shellfish acquire *Campylobacter* spp., and usually *C. jejuni*, from filter-feeding in water contaminated with fecal material [2]. The incidence of foodborne *Campylobacter* disease in the USA is 13.02 per 100,000, second only to Salmonella [18]. However, seafood-borne disease is rare [2, 69].

Staphylococcus

The genus *Staphylococcus* is made up of at least 40 species of gram-positive, facultative anaerobic organisms that are found throughout the world. Most of these organisms exist as commensal colonizing organisms of animals and humans, but may also be found in soil, on surfaces and in untreated water. They are hardy organisms that grow in the presence of bile salts and NaCl (up to 6.5%), and they can survive on many types of surfaces for extended periods of time making them a challenge to eliminate in public environments, such as gyms, prisons and hospitals. *Staphylococcal* species are differentiated from other important gram-positive organisms by the presence of the enzyme catalase, and they are differentiated amongst themselves by the presence of the enzyme coagulase which is present in the more clinically relevant organisms.

Among the organisms that make up this genus are a large variety of coagulase-negative staphylococci (CoNS) that are the primary commensal colonizers of humans. CoNS are pathogenic primarily for compromised populations only, such as preterm infants or persons with implanted prosthetic devices. *Staphylococcus aureus* is coagulase positive, and this species is the principle pathogen associated with human infection. In humans, *S. aureus* strains are opportunistic pathogens that may colonize, without infecting, up to 40% of the population [24, 79, 146], but may occasionally gain access to the host, evade the immune response, and causes disease [52, 53]. In addition, these organisms have acquired resistance to most of the antibiotics used against them making treatment of infections challenging.

Most infections caused by *S. aureus* are limited to the cutaneous tissues and are caused by a person's own colonizing organisms. However, these bacteria are also capable of causing serious, life threatening systemic disease. In fact, *S. aureus* including the methicillin resistant *S. aureus*, MRSA, are among the leading causes of *nosocomial* infections [10, 78]. Furthermore, MRSA have emerged as significant causes of community as well as hospital-associated infections [38, 78].

The diseases caused by *S. aureus* in humans and animals are often produced through the action of specific toxins or virulence factors that different bacterial isolates can produce. Specific toxins are associated with particular syndromes, such as toxic shock syndrome toxin and scalded skin syndrome. Many *S. aureus* are also capable of producing and secreting toxins responsible for staphylococcal food poisoning, termed enterotoxins, that only need to be ingested to cause intoxication and do not require the continued presence of the bacteria for disease. Consumption of seafood contaminated with *S. aureus* producing enterotoxins leads to staphylococcal food poisoning. Contamination of the food products often comes during processing as is seen in Listeriosis; however, organisms in the water and associated with marine life as seen in Peter the Great Bay, Japan, and Nha Trang Bay, South China, seas may also contribute to human disease [5].

S. aureus including drug resistant MRSA and CoNS have been isolated at recreational beaches, from marine ccc and temperate environments [5, 19, 50, 57, 133]. Adults and toddlers in diapers have been shown to shed *S. aureus* and the indicator organism *Enterococcus* into recreational marine waters and sand [118]. Persons using these recreational beaches may transmit and receive these organisms from the environment [42, 51, 131]. A retrospective epidemiological/microbiological monitoring study performed in Hawaii in subtropical marine waters and beaches found that persons were four times more likely to have staphylococcal skin infections if they had a history of seawater contact [20]. Whether there is a correlation of the microbial load in these environments and increased infections is yet to be verified; however, a recent study performed at a South Florida recreational beach did show a correlation between the average number of bathers in the water and the presence of *S. aureus* [119].

S. aureus and MRSA have also been isolated from marine mammals, including bottlenose dolphins (*Tursiops truncates*), seals, and walruses (blowholes, gastric fluids, fecal and anal cultures), both in captivity and in the wild, and have been associated with both colonization and disease [45, 105, 129]. In marine mammals in captivity, it is likely that the source of *S. aureus* and MRSA are colonized human handlers. The source of these organisms for the non-captive animals is not clear; however, colonized wild mammals were primarily identified in locations associated with human recreational use in the estuarine waters of Charleston, SC and Indian River Lagoon, FL [100, 129]. It has been suggested that some of the marine mammals, such as bottlenose dolphins, might serve as sentinels for transfer of resistant organisms from humans and animals into this environment, or simply indicate that the antibiotics are reaching this environment. To date, there are no confirmed cases of human infection from colonized or infected marine mammals.

Staphylococcus aureus and MRSA at Recreational Beaches

Since the early 1990s, investigators in Hawaii have isolated *S. aureus* from the waters used for recreation and suspected that exposure of bathers to the organisms in this environment might put them at risk for staphylococcal infections. Recent investigations have also isolated *S. aureus* and methicillin resistant, MRSA, from recreational marine beaches at multiple locations, including Hawaii, Puget Sound, California, and South Florida. The sources of the bacteria in these environments are likely multiple and not yet completely appreciated. *S. aureus* are not known to have a marine reservoir; however, humans and other mammals that may be present at the beach are known to be colonized with the bacteria and are potential sources. In fact, humans have been shown to readily shed their colonizing methicillin sensitive *S. aureus* (MSSA) and MRSA into marine waters and sand.

A recent study completed at a South Florida recreational beach collected MSSA and MRSA from ambient water, water nearby bathers and sand and evaluated the bacteria present in these environments for their potential to be associated with infection by determining the virulence factors they could produce. The study showed that 30–37% of water samples had *S. aureus*; however, it was unable to show a correlation between exposure to *S. aureus* and reported illness. The majority, greater than 97%, of *S. aureus* found at this location were MSSA that carried few virulence factors known to be associated with infection. However; the MRSA isolated from this location were similar to the MRSA found in the community that are known to have the potential to cause serious infections. The lack of association between exposure to *S. aureus* and illness in this study, at this location, and the lack of an adequate number of participants (sample size) to establish an association with an organism that is present in only present 37% of samples overall is explained in part by the low percentage of potentially virulent MRSA. The populations and concentrations of MSSA and MRSA at other more crowded, recreational beaches would likely be as different as the human populations utilizing these beaches. Bathers exposed to greater numbers of more virulent organisms could be at increased risk for infections. Further studies are required to establish the true risk to bathers exposed to MSSA and especially MRSA in these settings.

General Treatment Principles

Many of the marine or ocean-dwelling microorganisms important in human disease are gram-negative bacteria. The diseases caused by these microorganisms run the spectrum, from septicemia (e.g., *V. vulnificus*), to gastroenteritis, wound infections, ear infections, and eye infections. It is not surprising that treatment is also varied from syndrome to syndrome.

Cholera (*V. cholera*) is the classic example of a cause of noninflammatory severe gastroenteritis (rice water stools). In cholera, the diarrhea is toxin induced. The pathogenesis of diarrhea in other organisms may be inflammatory (e.g., *Salmonella, Campylobacter* or *Shigella*). The nature or mechanism of the diarrhea may affect the primary choice of, or decision regarding, treatment.

As a general rule, severe diarrheal diseases require fluid replacement, given either by mouth (oral rehydration) or by intravenous means as the primary mode of treatment, while inflammatory diarrheas require, in addition to fluids, also antimicrobial therapy. The mode of fluid delivery will depend on how sick the patient is, the availability of which treatment, and the capacity or resources available to deliver the fluid replacement to the patient.

Sometimes, antimicrobial treatment may be detrimental in severe diarrhea. A good example is the bloody diarrhea caused by *E. coli* 0157:H7, which may lead to hemolytic-uremic syndrome, especially in children exposed to both the toxin-producing *E. coli*, as well as antibiotic therapy [156]. In such cases, withdrawal of antimicrobial therapy may sometimes help prevent further complications of renal failure [108]. Treatment in this circumstance is entirely supportive.

For toxin-producing diarrheas, antibody binding in situ presents an attractive and elegant option, but this type of therapy has not yet been developed (for all practical purposes) for most diarrheal diseases seen in the clinical setting [101].

Septicemia (e.g., due to *V. vulnificus*) requires aggressive management in the intensive care unit (ICU) setting, surgical debridement where necessary, as well as antimicrobial therapy. Severe wound infections like necrotizing fasciitis require primary surgical debridement, antimicrobial treatment, and sometimes hyperbaric oxygen therapy.

Most patients with severe wound infections or septicemia do not typically die or have other adverse outcomes due to or as a consequence of antimicrobial resistance, but complications result because they present too late to the hospital, have devitalized tissues that were not promptly debrided, or for other host factor reasons.

In our experience, most *Vibrio* organisms seen in clinical practice on the Gulf Coast are sensitive to the third-generation cephalosporins (sometimes also to second generation cephalosporins), quinolones, tetracyclines, and aminoglycosides. The same goes for *Aeromonas hydrophila*, with similar susceptibilities as above, in addition to usual sensitivity to trimethorprim/sulfamethoxazole.

Acute diarrheal illnesses require antimicrobial therapy usually only for a short period (5 days is typically enough); while septicemia (e.g., typhoid fever and *Vibrio* sepsis) would require longer therapy (2 weeks or longer), depending on the complications.

Severe wound infections of the necrotizing fasciitis type often require multiple surgical debridements, in addition to antimicrobial therapy and local wound care. In addition, skin grafting or plastic surgery is often required to cover defective skin or tissue.

The key to successful treatment of all of these disease entities is timeliness in starting treatment. The earlier appropriate therapy is started, the better the clinical outcome. The later treatment is started, the more complications one can expect. Adjustments in antimicrobial therapy can and should be made after in vitro antimicrobial susceptibility studies are available. However, prompt treatment must be started very early, empirically (best guess or educated guess), before the laboratory susceptibility reports are available.

The Vibrios have predictable antimicrobial susceptibilities, more so than *Salmonella, E. coli*, or *Shigella*. Antimicrobial resistance to enteric pathogens reflects the pattern of use of antibiotics in a given environment, as well as the ease of antimicrobial drug availability and abuse in the area or locale where the infection was acquired. Typically for *Aeromonas*, resistance to ampicillin-like agents and first-generation cephalosporins is common; these agents should therefore not be used to treat infections due to *Aeromonas*.

In general, a gram-negative bacillus found in coastal or ocean water (outside and far away from a sewage drainage site) is likely to be free-living and, therefore, is more likely to be sensitive to multiple antimicrobial agents. Organisms causing disease acquired through human-to-human or foodborne transmission (e.g., *Salmonella*), on the other hand, are more likely to have been previously exposed to antibiotics (e.g., in animals or food products). Infections acquired through such contacts may therefore be more resistant to antimicrobials than free-living ocean, river, or brackish water bacteria [4].

The sensitivity of human-to-human or animal-to-human transmitted gram-negative bacteria (*E. coli, Salmonella, Shigella, Pseudomonas*, etc.) will usually reflect the pattern of local prevailing antimicrobial use in that community. The local hospital's antibiogram should provide the initial guide in the choice of empiric therapy, with necessary adjustments made after in vitro susceptibility studies are available.

Antimicrobial therapy for Brucellosis is often prolonged (up to 6 weeks) in order to prevent relapse. Often, combination therapy that includes a tetracycline plus rifampin, or an aminoglycoside, is required and recommended for complete cure of this debilitating disease [127, 132].

Acute gastroenteritis caused by food poisoning (e.g., staphylococcal preformed heat-stable enterotoxin) is often rapid in onset (within 1–6 h of food ingestion) and is also self-limited. The symptoms of severe nausea and vomiting occur usually within 1–6 h, and are usually over in less than 24 h [13]. Antimicrobial therapy is usually not required.

Treatment Considerations for Gram-Positive Organisms

Enterococci lack significant virulence factors associated with disease but are intrinsically resistant to many antibiotics currently in use. These bacteria are known to easily acquire antibiotic resistance genes from other microorganisms encountered in their environment. Treatment is guided by the determined antibiotic sensitivities of the infecting organisms and may be prolonged.

Streptococcus iniae infections usually present after exposure to fish with fever and cellulitis, often with bacterimia, and can be treated with intravenous penicillin and gentamicin [84]. *S. agalactiae* (Lancefield group B) remains sensitive to penicillin.

Listeria infections can be treated with common antibiotics, such as ampicillin, ciprofloxacin, linezolid and azithromicin. The delivery method of antibiotic is determined by the severity of diease.

Treatment for *Mycobacterium marinum* is driven by the severity of the infection [122] and ranges from oral monotherapy with minocycline, clarithromycin, doxycycline, ciprofloxacin, and trimethoprim-sulfamethoxazole for superficial cutaneous infections with susceptible organisms to combination therapies for drug-resistant strains. Severe infections, including those with a sporotrichoid distribution pattern, generally require combination therapy with rifampicin and ethambutol. Surgical debridement is not usually recommended however other alternative topical therapies such as cryotherapy, X-ray therapy, electrodesiccation, photodynamic therapy, and local hyperthermic therapy can be effective.

The majority of infections by *Staphylococcus aureus* are cutaneous infections limited to skin and soft tissues. Minor skin infections are usually treated with topical antibiotics, such as a nonprescription triple-antibiotic mixture or mupirocin. In some cases, oral antibiotics may be given for more severe skin infections. If abscesses are present, surgical drainage may be required and for smaller abscesses may be curative. More serious and life-threatening systemic infections are treated with intravenous antibiotics. The choice of antibiotic depends on the susceptibility of the particular staphylococcal strains, as determined by culture results in the laboratory. MRSA from the community may be sensitive to several antibiotics effective against MSSA; however, hospital-associated MRSA are usually resistant multiple antibiotics and may be challenging to treat. Vancomycin remains the drug of choice for multidrug resistant MRSA.

Future Directions

Clearly, most of what is known about waterborne human pathogens is based on laboratory and clinical observations, and this is especially true for autochothonous bacteria. The diseases and metabolic capabilities of these pathogens can be described in great detail, and in many cases the genomes of these bacteria have been completely sequenced. However, until their role or niche in the ocean and in freshwater habitats is fully investigated, it will not be possible to fully understand their ability to cause disease in humans. Accordingly, scientists need to continue asking the question, "how does this microorganism live in the ocean and yet invade humans to cause disease?"

Bibliography

Primary Literature

1. Abdelzaher AM, Wright ME, Ortega C, Solo-Gabriele HM, Miller G, Elmir S, Newman X, Shih P, Bonilla JA, Bonilla TD, Palmer CJ, Scott T, Lukasik J, Harwood VJ, McQuaig S, Sinigalliano C, Gidley M, Plano LR, Zhu X, Wang JD, Fleming LE (2010) Presence of pathogens and indicator microbes at a non-point source subtropical recreational marine beach. Appl Environ Microbiol 76:724–732
2. Abeyta C, Deeter FG, Kaysner CA, Stott RF, Wekell MM (1993) *Campylobacter jejuni* in a Washington state shellfish growing bed associated with illness. J Food Prot 56:323–325
3. Alonso JL, Alonso MA (1993) Presence of *Campylobacter* in marine waters of Valencia, Spain. Water Res 27:1559–1562
4. Baya AM, Brayton PR, Brown VL, Grimes DJ, Russek-Cohen E, Colwell RR (1986) Coincident plasmids and antimicrobial resistance in marine bacteria isolated from polluted and unpolluted Atlantic Ocean samples. Appl Environ Microbiol 51:1285–1292
5. Beleneva IA (2011) Incidence and characteristics of Staphylococcus aureus and Listeria monocytogenes from the Japan and South China seas. Mar Pollut Bull 62(2):382–387
6. Boerlin P, Boerlin-Petzold F et al (1997) Typing Listeria monocytogenes isolates from fish products and human listeriosis cases. Appl Environ Microbiol 63(4):1338–1343
7. Bortolussi R (2008) Listeriosis: a primer. CMAJ 179(8):795–797
8. Brock TD (1999) Robert Koch, a life in medicine and bacteriology. ASM Press, Washington, DC
9. Castro-Escarpulli G, Figuerasb MJ, Aguilera-Arreolaa G, Solerb L, Ferna´ndez-Rendo´na E, Aparicioa GO, Guarrob J, Chaco´n MR (2003) Characterisation of *Aeromonas* spp. isolated from frozen fish intended for human consumption in Mexico. Int J Food Microbiol 84:41–49
10. Centers for Disease Control and Prevention (1999) National Nosocomial Infections Surveillance system report: data from 1997–1999, Atlanta, GA
11. CDC (1996) Invasive infection due to *Streptococcus iniae* – Ontario, 1995–1996. Morb Mortal Wkly Rep 45:650–653
12. CDC (1998) Outbreak of *Vibrio parahaemolyticus* infections associated with eating raw oysters – Pacific Northwest, 1997. MMWR 47:457–462
13. CDC (2001) Diagnosis and management of foodborne illnesses: a primer for physicians. MMWR 50(RR02):1–69
14. CDC (2005) Rapid health response, assessment, and surveillance after a tsunami – Thailand, 2004–2005. MMWR 54:61–64
15. CDCb (2006) Vibrio parahaemolyticus infections associated with consumption of raw shellfish – three states, 2006. MMWR 55:1–2
16. CDCa (2006) Two cases of toxigentic Vibrio cholerage O1 infection after Hurricanes Katrina and Rita – Louisiana, October 2005. MMWR 55(02):31–32
17. CDC (2009) Preliminary FoodNet data on the incidence of infection with pathogens transmitted commonly through food – 10 states, 2008. MMWR 58:333–337
18. CDC (2010) Preliminary foodnet data on the incidence of infection with pathogens transmitted commonly through food – 10 states, 2009. MMWR 59:418–422
19. Charoenca N, Fujioka R (1993) Assessment of staphylococcus bacteria in Hawaii marine recreational waters. Water Sci Technol 27:283–289
20. Charoenca N, Fujioka RS (1995) Association of staphylococcal skin infections and swimming. Water Sci Technol 31:11–17
21. Chen C-Y, Wu K-M, Chang Y-C, Chang C-H, Tsai H-C, Liao T-L, Liu Y-M, Chen H-J, Shen AB-T, Li J-C et al (2003) Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. Cold Spring Harbor Laboratory Press, pp 2577–2587
22. Colwell RR, West PA, Maneval D, Remmers EF, Elliot EL, Carlson NE (1984) Ecology of the pathogenic vibrios in Chesapeake Bay. In: Colwell RR (ed) Vibrios in the environment. Wiley, New York
23. Chowdhury NR, Chakraborty S, Ramamurthy T, Nishibuchi M, Yamasaki S, Takeda Y, Nair GB (2000) Molecular evidence of clonal Vibrio parahaemolyticus pandemic strains. Emerg Infect Dis 6:631–636
24. Cole AM, Tahk S, Oren A, Yoshioka D, Kim YH, Park A, Ganz T (2001) Determinants of *Staphylococcus aureus* nasal carriage. Clin Diagn Lab Immunol 8:1064–1069
25. Colwell RR (1996) Global climate and infectious disease: the cholera paradigm. Science 274:2025–2031
26. Colwell RR, Huq A, Islam MS, Aziz KMA, Yunus M, Khan NH, Mahmud A, Sack RB, Nair GB, Chakraborty J, Sack DA, Russek-Cohen E (2003) Reduction of cholera in Bangladeshi villages by simple filtration. Proc Natl Acad Sci 100:1051–1055

27. Colwell RR (2006) A global and historical perspective of the genus *Vibrio*. In: Thompson FL, Austin B, Swings J (eds) The biology of Vibrios. ASM Press, Washington, DC, pp 3–11
28. Constantin de Magny G, Murtugudde R, Sapiano MRP, Nizam A, Brown CW, Busalacchi AJ, Yunus M, Nair GB, Gil AI, Lanata CF, Calkins J, Manna B, Rajendran K, Bhattacharya MK, Huq A, Sack RB, Colwell RR (2008) Environmental signatures associated with cholera epidemics. Proc Natl Acad Sci 105:17676–17681
29. Cook DW, Ruple AD (1989) Indicator bacteria and Vibrionaceae multiplication in post-harvest shellstock oysters. J Food Prot 52:343–349
30. Croci L, Suffredini E, Cozzi L, Paniconi M, Ciccaglioni G, Colombo MM (2007) Evaluation of different polymerase chain reaction methods for the identification of *Vibrio parahaemolyticus* strains isolated by cultural methods. J AOAC Int 90:1588–1597
31. Dalgaard P, Vancanneyt M et al (2003) Identification of lactic acid bacteria from spoilage associations of cooked and brined shrimps stored under modified atmosphere between 0 degrees C and 25 degrees C. J Appl Microbiol 94(1):80–89
32. DePaola A, Kaysner CA, Bowers J, Cook DW (2000) Environmental investigations of *Vibrio parahaemolyticus* in oysters after outbreaks in Washington, Texas, and New York (1997 and 1998). Appl Environ Microbiol 66:4649–4654
33. DePaola A, Nordstrom JL, Bowers JC, Wells JG, Cook DW (2003) Seasonal abundance of total and pathogenic *Vibrio parahaemolyticus* in Alabama oysters. Appl Environ Microbiol 69:1521–1526
34. DePaola A, Jones JL, Woods J, Burkhardt W III, Calci KR, Krantz JA, Bowers JC, Kasturi K, Byars RH, Jacobs E, Williams-Hill D, Nabe K (2010) Bacterial and viral pathogens in live oysters: 2007 United States survey. Appl Environ Microbiol 76:2754–2768
35. DeLong EF, Pace NR (2000) Environmental diversity of bacteria and archaea. Syst Biol 50:470–478
36. Destro MT (2000) Incidence and significance of Listeria in fish and fish products from Latin America. Int J Food Microbiol 62(3):191–196
37. Dibrov P (2005) The sodium cycle in *Vibrio cholerae*: riddles in the dark. Biochemistry (Mosc) 70:150–153
38. Diep BA, Carleton HA et al (2006) Roles of 34 virulence genes in the evolution of hospital- and community-associated strains of methicillin-resistant *Staphylococcus aureus*. J Infect Dis 193(11):1495–1503
39. Eckburg PB, Lepp PW, Relman DA (2003) Archaea and their potential role in human disease. Infect Immun 71:591–596
40. Edelstein H (1994) Mycobacterium marinum skin infections. Report of 31 cases and review of the literature. Arch Intern Med 154(12):1359–1364
41. Eldar A, Bejerano Y, Bercovier H (1994) *Streptococcus shiloi* and *Streptococcus difficile*: two new streptococcal species causing a meningoencephalitis in fish. Curr Microbiol 28:139–143
42. Elmir SM, Wright ME, Abdelzaher A, Solo-Gabriele HM, Fleming LE, Miller G, Rybolowik M, Peter Shih MT, Pillai SP, Cooper JA, Quaye EA (2007) Quantitative evaluation of bacteria released by bathers in a marine water. Water Res 41:3–10
43. Elmir SM, Shibata T, Solo-Gabriele HM, Sinigalliano CD, Gidley ML, Miller G, Plano LR, Kish J, Withum K, Fleming LE (2009) Quantitative evaluation of enterococci and Bacteroidales released by adults and toddlers in marine water. Water Res 43:4610–4616
44. FAO (2010) FAO expert workshop on the application of biosecurity measures to control *Salmonella* contamination in sustainable aquaculture. Food and Agriculture Organization of the United Nations, Rome (ISBN 978-92-5-106553-2)
45. Faires MC, Gehring E et al (2009) Methicillin-resistant Staphylococcus aureus in marine mammals. Emerg Infect Dis 15(12):2071–2072
46. Farmer JJ (1980) Revival of the name *Vibrio vulnificus*. Int J Syst Bacteriol 30:656
47. Farmer JJI, Janda M, Brenner FW, Cameron DN, Birkhead KM (2005) Genus 1. *Vibrio*. In: Brenner DJ, Krieg NR, Staley JT, Garrity GM (eds) Bergey's manual of systematic bacteriology, vol 2, The proteobacteria, Part B. The gammaproteobacteria. Springer, New York, pp 494–546
48. FDA (2009) Bad bugs book: foodborne pathogenic microorganisms and natural toxins handbook. http://www.fda.gov/Food/FoodSafety/FoodborneIllness/FoodborneIllnessFoodbornePathogensNaturalToxins/BadBugBook/ucm070523.htm
49. Fisher K, Phillips C (2009) The ecology, epidemiology and virulence of Enterococcus. Microbiology 155(Pt 6): 1749–1757
50. Fleisher JM, Fleming LE et al (2010) The BEACHES Study: health effects and exposures from non-point source microbial contaminants in subtropical recreational marine waters. Int J Epidemiol 39(5):1291–1298
51. Fleming L, Solo Gabriel H et al (2008) Final report on the pilot epidemiologic assessment of microbial indicators for monitoring recreational water quality in marine sub/tropical environments. The NSF NIEHS OHH Center, Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami
52. Foster TJ (2004) The *Staphylococcus aureus* "superbug". J Clin Invest 114(12):1693–1696
53. Foster TJ (2005) Immune evasion by staphylococci. Nat Rev Microbiol 3(12):948–958
54. Fujino T, Okuno Y, Nakada D, Aoyama A, Fukai K, Mukai T, Ueho T (1953) On the bacteriological examination of shirasu-food poisoning. Med J Osaka Univ 4:299–304
55. Gefen O, Mumcuoglu M, Engel G, Balaban NQ (2008) Single-cell protein induction dynamics reveals a period of vulnerability to antibiotics in persister bacteria. Proc Natl Acad Sci 105:6145–6149
56. Gil AL, Louis VR, Rivera ING, Lipp E, Huq A, Lanata CF, Taylor DN, Russek-Cohen E. Choopun N, Sack RB, Colwell RR (2004) Occurance and distribution of *Vibrio cholerae* in the coastal environment of Peru. Environ Microbiol 6:699–706
57. Goodwin KD, Pobuda M (2009) Performance of CHROMagar Staph aureus and CHROMagar MRSA for detection of Staphylococcus aureus in seawater and beach sand–comparison of culture, agglutination, and molecular analyses. Water Res 43:4802–4811

58. Grimes DJ, Singleton FL, Stemmler J, Palmer LM, Brayton P, Colwell RR (1984) Microbiological effects of wastewater effluent discharge into coastal waters of Puerto Rico. Water Res 18:613–619
59. Grimes DJ, Mills AL, Nealson KH (2000) The importance of viable but nonculturable bacteria in biogeochemistry. In: Colwell RR, Grimes DJ (eds) Nonculturable microorganisms in the environment. ASM Press, Washington, DC, pp 209–227
60. Guvener ZT, McCarter LL (2003) Multiple regulators control capsular polysaccharide production in *Vibrio parahaemolyticus*. J Bacteriol 185:5431–5441
61. Hara-Kudo Y, Saito S, Ohtsuka K, Yamasaki S, Yahiro S, Nishio T, Iwade Y, Otomo T, Konuma H, Tanaka H, Nakagawa H, Sugiyama K, Sugita-Konishi Y, Kumagai S (2010) Decreasing *Vibrio parahaemolyticus* infections and analysis of seafood contamination in Japan. Vibrios in the Environment-2010, Biloxi, MS (Abstracts)
62. Harth E, Matsuda L, Hernández C, Rioseco ML, Romero J, González-Escalona N, Martínez-Urtaza J, Espejo RT (2009) Epidemiology of *Vibrio parahaemolyticus* outbreaks, southern Chile. Emerg Infect Dis 15:163–168
63. Harwood VJ, Whitlock J, Withington V (2000) Classification of antibiotic resistance patterns of indicator bacteria by discriminant analysis: use in predicting the source of fecal contamination in subtropical waters. Appl Environ Microbiol 66: 3698–3704
64. He H, Wang Q, Sheng L, Liu Q, Zhang Y (2010) Functional characterization of *Vibrio alginolyticus* twin-arginine translocation system: Its roles in biofilm formation, extracellular protease activity, and virulence towards fish. Curr Microbiol. doi:10.1007/s00284-010-9844-6
65. Hensel M (2004) Evolution of pathogenicity islands of *Salmonella enterica*. Int J Med Microbiol 294:95–102
66. Ho PL, Tang WM, Lo KS, Yuen KY (1998) Necrotizing fasciitis due to *Vibrio alginolyticus* following an injury inflicted by a stingray. Scand J Infect Dis 30:192–193
67. Honda T, Iida T (1993) The pathogenicity of *Vibrio parahaemolyticus* and the role of thermostable direct haemolysin and related haemolysins. Rev Med Microbiol 4:106–113
68. Huq A, Sack RB, Nizam A, Longini IM, Balakrish Nair G, Ali A, Morris JG Jr, Khan MNH, Siddique AK, Yunus M, Albert MJ, Sack DA, Colwell RR (2005) Critical factors influencing the occurrence of *Vibrio cholerae* in the environment of Bangladesh. Appl Environ Microbiol 71:4645–4654
69. Jacobs-Reitsma W (2000) *Campylobacter* in the food supply. In: Nachamkin I, Blaser MJ (eds) Campylobacter, 2nd edn. ASM Press, Washington, DC, pp 467–480
70. Jaffres E, Sohier D et al (2009) Study of the bacterial ecosystem in tropical cooked and peeled shrimps using a polyphasic approach. Int J Food Microbiol 131(1):20–29
71. Jensen AE, Cheville NF, Thoen CO, MacMillan AP, Miller WG (1999) Genomic fingerprinting and development of a dendrogram for *Brucella* spp. isolated from seals, porpoises, and dolphins. J Vet Diagn Invest 11:152–157
72. Jia A, Woo NY, Zhang XH (2010) Expression, purification, and characterization of thermolabile hemolysin (TLH) from *Vibrio alginolyticus*. Dis Aquat Organ 90:121–127
73. Johnson CN, Flowers AR, Young VC, Gonzalez-Escalona N, DePaola A, Grimes DJ (2009) Genetic relatedness among *tdh*$^+$ and *trh*$^+$ *Vibrio parahaemolyticus* cultured from Gulf of Mexico oysters (*Crassostrea virginica*) and surrounding water and sediment. Microb Ecol 57:437–443
74. Johnson CN, Flowers AR, Noriea NF III, Zimmerman AM, Bowers J, DePaola A, Grimes DJ (2010) Relationships between environmental factors and pathogenic vibrios in the northern Gulf of Mexico. Appl Environ Microbiol 76:7076–7084
75. Joseph SW, Daily OP, Hunt WS, Seidler RJ, Allen DA, Colwell RR (1979) *Aeromonas* primary wound infection of a diver in polluted waters. J Clin Microbiol 10:46–49
76. Kaper JB, Morris JG, Levine MM (1995) Cholera. Clin Microbiol Rev 8:48–86
77. Kitao T, Aoki T, Sakoh R (1981) Epizootic caused by beta-hemolytic *Streptococcus* species in cultured freshwater fish. Fish Pathol 19:173–180
78. Klevens RM, Morrison MA, Nadle J, Petit S, Gershman K, Ray S, Harrison LH, Lynfield R, Dumyati G, Townes JM, Craig AS, Zell ER, Fosheim GE, McDougal LK, Carey RB, Fridkin SK (2007) Invasive methicillin-resistant Staphylococcus aureus infections in the United States. JAMA 298:1763–1771
79. Kluytmans J, van Belkum A, Verbrugh H (1997) Nasal carriage of *Staphylococcus aureus*: epidemiology, underlying mechanisms, and associated risks. Clin Microbiol Rev 10:505–520
80. Kusuda R (1992) Bacterial fish diseases in marineculture in Japan with special emphasis on streptococcosis. Isr J Aquacult 44:140
81. Lappi VR, Ho A et al (2004) Prevalence and growth of Listeria on naturally contaminated smoked salmon over 28 days of storage at 4 degrees C. J Food Prot 67(5):1022–1026
82. Lappi VR, Thimothe J et al (2004) Longitudinal studies on Listeria in smoked fish plants: impact of intervention strategies on contamination patterns. J Food Prot 67(11):2500–2514
83. Lappi VR, Thimothe J et al (2004) Impact of intervention strategies on Listeria contamination patterns in crawfish processing plants: a longitudinal study. J Food Prot 67(6):1163–1169
84. Lau SK, Woo PC et al (2003) Invasive Streptococcus iniae infections outside North America. J Clin Microbiol 41(3):1004–1009
85. Lauková A, Juris P (1997) Distribution and characterization of *Enterococcus* species in municipal sewages. Microbios 89:73–80
86. Li XC, Xiang ZY, Xu XM, Yan WH, Ma JM (2009) Endophthalmitis Caused by *Vibrio alginolyticus*. J Clin Microbiol 47:3379–3381
87. Lianou A, Sofos JN (2007) A review of the incidence and transmission of Listeria monocytogenes in ready-to-eat products in retail and food service environments. J Food Prot 70(9):2172–2198

88. Linell F, Norden A (1954) Mycobacterium balnei, a new acid-fast bacillus occurring in swimming pools and capable of producing skin lesions in humans. Acta Tuberc Scand Suppl 33:1–84
89. Liu D (ed) (2008) Handbook of *Listeria monocytogenes*. CRC Press, Boca Raton
90. Louis VR, Russek-Cohen E, Choopun N, Rivera ING, Gangle B, Jiang SC, Rubin A, Patz JA, Huq A, Colwell RR (2003) Predictability of *Vibrio cholerae* in Chesapeake Bay. Appl Environ Microbiol 69:2773–2785
91. Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, Tagomori K, Iijima Y, Najima M, Nakano M, Yamashita A, Kubota Y, Kimura S, Yasunaga T, Honda T, Shinagawa H, Hattori M, Iida T (2003) Genome sequence of *Vibrio parahaemolyticus* a pathogenic mechanism distinct from that of *V. cholerae*. Lancet 361:743–749
92. Manero A, Blanch AR (1999) Identification of *Enterococcus* spp. with a biochemical key. Appl Environ Microbiol 65:4425–4430
93. Maquart M, LeFleche P, Foster G, Tryland M, Ramisse F, Djønne B, Al Dahouk S, Jacques I, Neubauer H, Walravens K, Godfroid J, Cloeckaert A, Vergnaud G (2009) MLVA-16 typing of 295 marine mammal *Brucella* isolates from different animal and geographic origins identifies 7 major groups within *Brucella ceti* and *Brucella pinnipedialis*. BMC Microbiol 9:145. doi:10.1186/1471-2180-9-145
94. McCarter LL (2006) Motility and chemotaxis. In: Thompson FL, Austin B, Swings J (eds) The biology of Vibrios. ASM Press, Washington, DC, pp 115–132
95. McDonald WL, Jamaludin R, Mackereth G, Hansen M, Humphrey S, Short P, Taylor T, Swingler J, Dawson CE, Whatmore AM, Stubberfield E, Perrett LL, Simmons G (2006) Characterisation of a *Brucella* sp. strain as a marine-mammal type despite isolation from a patient with spinal osteomyelitis in New Zealand. J Clin Microbiol 44:4363–4370
96. McLaughlin JB, DePaola A, Bopp CA, Martinek KA, Napolilli NP, Allison CG, Murray SL, Thompson EC, Bird MM, Middaugh JP (2005) Outbreak of *Vibrio parahaemolyticus* gastroenteritis associated with Alaskan oysters. N Engl J Med 353:1463–1470
97. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe RV (1999) Food-related illness and death in the United States. Emerg Infect Dis 5:607–625
98. Meador CE, Parsons MM, Bopp CA, Gerner-Smidt P, Painter JA, Vora GJ (2007) Virulence gene- and pandemic group-specific marker profiling of clinical *Vibrio parahaemolyticus* isolates. J Clin Microbiol 45:1133–1139
99. Mejlholm O, Kjeldgaard J et al (2008) Microbial changes and growth of Listeria monocytogenes during chilled storage of brined shrimp (Pandalus borealis). Int J Food Microbiol 124(3):250–259
100. Morris PJ, Johnson WR et al (2011) Isolation of culturable microorganisms from free-ranging bottlenose dolphins (*Tursiops truncatus*) from the southeastern United States. Vet Microbiol 148(2–4):440–447
101. Mukherjee J, Chios K, Fishwild D, Hudson D, O'Donnell S, Rich SM, Donohue-Rolf A, Tzipori S (2002) Human Stx2-specific monoclonal antibodies prevent systemic complications of *Escherichia coli* O157:H7 infection. Infect Immun 70(2): 612–619
102. Nishibuchi M (2006) Miscellaneous human pathogens. In: Thompson FL, Austin B, Swings J (eds) The biology of Vibrios. ASM Press, Washington, DC, pp 367–381
103. Noriea NF III, Johnson CN, Griffitt KJ, Grimes DJ (2010) Distribution of type III secretion systems in *Vibrio parahaemolyticus* from the Northern Gulf of Mexico. J Appl Microbiol 109:953–962
104. Oliver JD (2006) *Vibrio vulnificus*. In: Thompson FL, Austin B, Swings J (eds) The biology of Vibrios. ASM Press, Washington, DC, pp 349–366
105. O'Mahony R, Abbott Y et al (2005) Methicillin-resistant Staphylococcus aureus (MRSA) isolated from animals and veterinary personnel in Ireland. Vet Microbiol 109(3–4): 285–296
106. O'Neill KR, Jones SH, Grimes DJ (1992) Seasonal incidence of *Vibrio vulnificus* in the Great Bay Estuary of New Hampshire and Maine. Appl Environ Microbiol 58:3257–3262
107. Oshiro R, Fujioka R (1995) Sand, soil, and pigeon droppings – sources of indicator bacteria in the waters of Hanauma Bay, Oahu, Hawaii. Water Sci Technol 31:251–254
108. Panos GZ, Betsi GI, Falagas ME (2006) Systematic review: are antibiotics detrimental or beneficial for the treatment of patients with *Escherichia coli* O157:H7 infection? Aliment Pharmacol Ther 24(5):731–742
109. Paranjpye RN, Strom MS (2005) A *Vibrio vulnificus* Type IV pilin contributes to biofilm formation, adherence to epithelial cells, and virulence. Infect Immun 73:1411–1422
110. Paranjpye RN, Johnson AB, Baxter AE, Strom MS (2007) Role of Type IV pilins in persistence of *Vibrio vulnificus* in *Crassostrea virginica* oysters. Appl Environ Microbiol 73:5041–5044
111. Parenti DM, Symington JS et al (1995) Mycobacterium kansasii bacteremia in patients infected with human immunodeficiency virus. Clin Infect Dis 21(4):1001–1003
112. Park K-S, Iida T, Yamaichi Y, Oyagi T, Yamamoto K, Honda T (2000) Genetic characterization of DNA region containing the *trh* and *ure* genes of *Vibrio parahaemolyticus*. Infect Immun 68:5742–5748
113. Peir GB, Madin SH (1976) *Streptococcus iniae* sp. nov., a beta-hemolytic streptococcus isolated from an Amazon freshwater dolphin, *Inia geoffrensis*. Int J Syst Bacteriol 26:545–553
114. Pellett S, Bigley DV, Grimes DJ (1983) Distribution of *Pseudomonas aeruginosa* in a riverine ecosystem. Appl Environ Microbiol 45:328–332
115. Perera RP, Johnson SK, Collins MD, Lewis DH (1994) *Streptococcus iniae* associated with mortality of *Tilapia nilotica* and *T. aurea* hybrids. J Aquat Anim Health 6:335–340
116. Petersen A, Dalsgaard A (2003) Species composition and antimicrobial resistance genes of Enterococcus spp, isolated from integrated and traditional fish farms in Thailand. Environ Microbiol 5(5):395–402

117. Phillips AMB, DePaolo A, Bowers J, Ladner S, Grimes DJ (2007) An evaluation of the use of remotely sensed parameters for prediction of incidence and risk associated with *Vibrio parahaemolyticus* in Gulf Coast oysters (*Crassostrea virginica*). J Food Prot 70:879–884
118. Plano LR, Garza AC et al (2011) Shedding of Staphylococcus aureus and methicillin-resistant Staphylococcus aureus from adult and pediatric bathers in marine waters. BMC Microbiol 11(1):5
119. Plano LRW, Shibata T et al (2011) Characterization of *Staphylococcus aureus* and community associated MRSA at a recreational marine beach in South Florida. J Antimicrob Chemother (in review)
120. Posfay-Barbe KM, Wald ER (2009) Listeriosis. Semin Fetal Neonatal Med 14(4):228–233
121. Qasem JA, Sameer A-Z, Salwa A-M, Samee A-A, Ahmed A-M, Al-Sharifi Faisal A-S (2008) Molecular investigation of *Streptococcus agalactiae* isolates from environmental samples and fish specimens during a massive fish kill in Kuwait Bay. Pakistan J Biol Sci 11:2500–2504
122. Rallis E, Koumantaki–Mathioudaki E (2007) Treatment of Mycobacterium marinum cutaneous infections. Expert Opin Pharmacother 8(17):2965–2978
123. Reichelt JL, Bauman P, Bauman L (1979) Study of genetic relationships among marine species *Beneckea* and *Photobacterium* by means of DNA/DNA hybridization. Arch Microbiol 1110:101–120
124. Rice EW, Messer JW et al (1995) Occurrence of high-level aminoglycoside resistance in environmental isolates of enterococci. Appl Environ Microbiol 61(1):374–376
125. Rivas AL, Gonzalez RN, Wiedmann M, Bruce JL, Cole EM et al (1997) Diversity of *Streptococcus agalactiae* and *Staphylococcus aureus* ribotypes recovered from New York dairy herds. Amer J Vet Res 58:482–487
126. Rivers B, Steck TR (2001) Viable but nonculturable uropathogenic bacteria are present in the mouse urinary tract following urinary tract infection and antibiotic therapy. Urol Res 29:60–66
127. Roushan MRH, Mohraz M, Hajiahmadi M, Ramzani A, Valayati AA (2006) Efficacy of Gentamicin plus Doxycycline versus Streptomycin plus Doxycycline in the treatment of Brucellosis in humans. Clin Infect Dis 42:1075–1080
128. Saubolle M (1989) Nontuberculous mycobacteria as agents in human disease in the United States. Clin Microbiol Newslett 11:113–117
129. Schaefer AM, Goldstein JD et al (2009) Antibiotic-resistant organisms cultured from Atlantic bottlenose dolphins (*Tursiops truncatus*) inhabiting estuarine waters of Charleston, SC and Indian River Lagoon, FL. Ecohealth 6(1):33–41
130. Shime-Hattori A et al (2006) Two type IV pili of *Vibrio parahaemolyticus* play different roles in biofilm formation. FEMS Microbiol Lett 264:89–97
131. Sinigalliano CD, Fleisher JM et al (2010) Traditional and molecular analyses for fecal indicator bacteria in non-point source subtropical recreational marine waters. Water Res 44(13):3763–3772
132. Skalsky K, Yahav D, Bishara J, Pitlik S, Leibovici L, Paul M (2008) Treatment of human brucellosis: systematic review and meta-analysis of randomized trials. BMJ 336(7646):701–704
133. Soge OO, Meschke JS, No DB, Roberts MC (2009) Characterization of methicillin resistant Staphylococcus aureus and methicillin-resistant coagulase negative Staphylococcus spp isolated from US West coast public marine beaches. J Antimicrob Chemother 64:1148–1155
134. Sun B, Zhang XH, Tang X, Wang S, Zhong Y, Chen J, Austin B (2007) A single residue change in *Vibrio harveyi* hemolysin results in the loss of phospholipase and hemolytic activities and pathogenicity for turbot (*Scophthalmus maximus*). J Bacteriol 189:2575–2579
135. Takeda T, Peina Y, Ogawa A, Dohi S, Abe H, Nair GB, Pal SC (1991) Detection of heat-stable enterotoxin in a cholera toxin gene-positive strain of *Vibrio cholerae* O1. FEMS Microbiol Lett 64:23–27
136. Tauxe R, Seminario L, Tapia R, Libel M (1994) The Latin American epidemic. In: Wachsmuth IK, Blake PA, Olsvik O (eds) *Vibrio cholerae* and cholera: molecular to global perspectives. American Society for Microbiology, Washington, DC
137. Thompson FL, Hoste B, Vandemeulebroecke K, Swings J (2003) Reclassification of *Vibrio hollisae* as *Grimontia hollisae* gen. nov., comb. Nov. Int J Syst Evol Microbiol 53:1615–1617
138. Tokuda H, Unemoto T (1982) Characterization of the respiration-dependant Na^+ pump in the marine bacterium *Vibrio alginolyticus*. J Biol Chem 257:10007–10014
139. Tsai YH, Kung H-F, Lee T-M, Lin G-T, Hwang D-F (2004) Histamine-related hygienic quantities and bacteria found in popular commercial scombroid fish fillets in Taiwan. J Food Prot 67:407–412
140. Uh Y, Park JS, Hwang GY, Jang IH, Yoon KJ, Park HC, Hwang SO (2001) *Vibrio alginolyticus* acute gastroenteritis: a report of two cases. Clin Microbiol Infect 7:104–106
141. Ulitzer S (1975) The mechanism of swarming of *Vibrio alginolyticus*. Arch Microbiol 104:67–71
142. USEPA (2000) Improved enumeration methods for the recreational water quality indicators: Enterococci and *Escherichia coli*. EPA/821/R-97/004, Washington, DC
143. Valdivia E, Martin-Sanchez I et al (1996) Incidence of antibiotic resistance and sex pheromone response among enterococci isolated from clinical human samples and from municipal waste water. J Appl Bacteriol 81(5):538–544
144. Vally H, Whittle A, Cameron S, Dowse GK, Watson T (2004) Outbreak of *Aeromonas hydrophila* wound infections associated with mud football. Clin Infect Dis 38:1084–1089
145. Vivekanandhana G, Hathab AAM, Lakshmanaperumalsamy P (2005) Prevalence of *Aeromonas hydrophila* in fish and prawns from the seafood market of Coimbatore, South India. Food Microbiol 22:133–137
146. Von Eiff C, Becker K, Machka K, Stammer H, Peters G (2001) Nasal carriage as a source of Staphylococcus aureus bacteremia. Study Group. N Engl J Med 344:11–16

147. Wade TJ, Pai N et al (2003) Do U.S. Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. Environ Health Perspect 111(8):1102–1109
148. Waldor MK, Mekalanos JJ (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 272:1910–1914
149. Wang SX, Zhang XH, Zhong YB, Sun BG, Chen JX (2007) Genes encoding the *Vibrio harveyi* haemolysin (VHH)/thermolabile haemolysin (TLH) are widespread in vibrios. Wei Sheng Wu Xue Bao 47:874–881
150. Weinstein MR, Litt M, Kertesz DA, Wyper P, Rose D, Coulter M, McGeer A, Facklam R, Ostach C, Willey BM, Borczyk A, Low DE et al (1997) Invasive infections due to a fish pathogen, Streptococcus iniae. N Engl J Med 337:589–594
151. Wendt JR, Lamm RC et al (1986) An unusually aggressive Mycobacterium marinum hand infection. J Hand Surg Am 11(5):753–755
152. Whatmore AM, Dawson CE, Groussaud P, Koylass MS, King AC, Shankster SJ, Sohn AH, Probert WS, McDonald WL (2008) Marine mammal *Brucella* genotype associated with zoonotic infection. Emerg Infect Dis 14:517–518
153. WHO (World Health Organization) (2010) Cholera, 2009. Wkly Epidemiol Rec 85:293–308
154. Wilson IG, McAfee GG (2002) Vancomycin-resistant enterococci in shellfish, unchlorinated waters, and chicken. Int J Food Microbiol 79(3):143–151
155. Wolinsky E (1992) Mycobacterial diseases other than tuberculosis. Clin Infect Dis 15(1):1–10
156. Wong CS, Jelacic S, Habeeb RL, Watkins SL, Tarr PI (2000) The risk of the hemolytic-uremic syndrome after antibiotic treatment of *Escherichia coli* O157:H7 infections. N Engl J Med 342(26):1930–1936
157. Yakushi T, Maki S, Homma M (2004) Interaction of PomB with the third transmembrane segment of PomA in the Na$^+$-driven polar flagellum of *Vibrio alginolyticus*. J Bacteriol 186:5281–5291
158. Yamahara KM, Walters SP et al (2009) Growth of enterococci in unaltered, unseeded beach sands subjected to tidal wetting. Appl Environ Microbiol 75(6):1517–1524
159. Zimmerman AM, Rebarchik DM, Flowers AR, Williams JL, Grimes DJ (2009) *Escherichia coli* detection using mTEC agar and fluorescent antibody direct viable counting on coastal recreational water samples. Lett Appl Microbiol 49:478–483

Books and Reviews

Belkin S, Colwell RR (eds) (2005) Pathogenic microorganisms in the sea. Kluwer/Plenum, New York

Colwell RR, Grimes DJ (eds) (2000) Nonculturable microorganisms in the environment. ASM Press, Washington, DC

Walsh PJ, Smith SL, Fleming LE, Solo-Gabriele HM, Gerwick WH (eds) (2008) Oceans and human health: risk and remedies from the sea. Elsevier, St. Louis

Hurricane and Monsoon Tracking with Driftsondes

PHILIPPE DROBINSKI[1], PHILIPPE COCQUEREZ[2],
A. DOERENBECHER[3], TERRENCE HOCK[4], C. LAVAYSSE[1],
D. PARSONS[4], J. L. REDELSPERGER[3]
[1]Institut Pierre Simon Laplace/Laboratoire de Météorologie Dynamique, École Polytechnique/UPMC/ENS/CNRS, Palaiseau, France
[2]Sous-direction Ballon, Centre National d'Études Spatiales, Toulouse, France
[3]Centre National de Recherches Météorologiques, Météo-France/CNRS, Toulouse, France
[4]Atmospheric Technology Division, National Center for Atmospheric Research, Boulder, CO, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Driftsonde Observations
Analysis of the Driftsonde Trajectories and the Respective Synoptic Environment
Analyses of AEW with the Driftsondes
Conclusion
Future Directions
Acknowledgments
Bibliography

Glossary

African easterly jet (AEJ) Wind jet developing at about 600-hPa pressure level (about 4,200 m) and 5°N latitude over the African continent.

African easterly waves (AEWs) Synoptic scale, westward propagating waves embedded in the AEJ.

AMMA African Monsoon Multidisciplinary Analysis (AMMA) program, initiated in 2002 with an intensive field campaign in 2006, and focusing on the physical changes in the environment of the West African monsoon and their impacts on society.

Boundary layer pressurized balloon (BLPB) Balloon drifting in the atmospheric boundary layer and carrying an instrumented gondola which collects measurements of the typical meteorological variables.

Concordiasi International project of the THORPEX-International Polar Year aiming at a better understanding of the climate of Antarctica.

Driftsonde Stratospheric drifting balloon launching dropsondes over high-impact weather, providing vertical profiles of meteorological data.

Hurricane A hurricane is a tropical cyclone, occurring in the North Atlantic Ocean or the Northeast Pacific Ocean. A tropical cyclone is a storm system characterized by a large low-pressure center and numerous thunderstorms that produce strong winds and heavy rain.

HyMeX Hydrological cycle in the Mediterranean experiment (HyMeX) program aiming at improving the understanding of running-water cycle in the Mediterranean region with a particular focus on the evolution and predictability of hydro-meteorological extreme events in the perspective of climate change.

Mesoscale convective system (MCS) Complex of thunderstorms normally persisting for several hours or more.

Saharan air layer (SAL) Surges of hot, dry air that cascade into the Atlantic from the Sahel region of Africa.

THORPEX Global program of the World Meteorological Organization (WMO) aiming at accelerating improvements in the prediction of high-impact weather.

T-NAWDEX THORPEX-North Atlantic Waveguide and Downstream impact Experiment.

T-PARC THORPEX Pacific Asian Regional Campaign.

West African monsoon Rainy phase of a seasonal change in atmospheric circulation over West Africa. The other major monsoon systems are the Asia-Australian monsoons in the sampled environment.

Definition of the Subject

Tropical cyclones (TCs) are a typical weather threat. The threat can apply to humans, their properties, and activities. Their prediction, particularly their trajectory and intensity, remains difficult. In addition, TCs develop above the tropical oceans where the coverage by in situ observations is poor and within cloud clusters (mesoscale convective systems MCS) that limit the ability of numerical weather prediction (NWP) models to assimilate satellite data [18]. Improved forecast of TCs trajectories is a huge benefit in terms of material costs of evacuations and damage, not being able to quantify saved life.

The deployment of additional observations to improve understanding and forecasting TCs started very early in the USA, from the early 1980s. This approach is called adaptive observation and was initiated by the National Oceanic and Atmospheric Administration (NOAA). The philosophy of adaptive observation is the adjustment of the usual (so-called routine) observing network in order to deal with some particularly threatening weather situations to come. This adaptation aims at improving the numerical weather prediction (NWP) through the deployment of additional observations. When incorporated in the models, thanks to data assimilation procedures, these added observations decrease the uncertainty on the modeled state of the atmosphere (analysis). Subsequent numerical weather forecasts are likely to be improved. This improvement is expressed with respect to the NWP system that would not incorporate the additional data.

Any type of observation can be used. All time and space scales may be considered. Historically, the adaptive observation focused on the synoptic scales, for TCs first, then for mid-latitudes winter storms [15, 22]. The terms "adaptive observation" or "observation targeting" may be used in various senses that may be misleading. Adaptive observation first refers to the observing strategy itself, which is a real-time observation practice in close link with NWP.

An adaptive observing strategy requires a constant monitoring of the weather situation. The typical strategy is implemented following five stages:

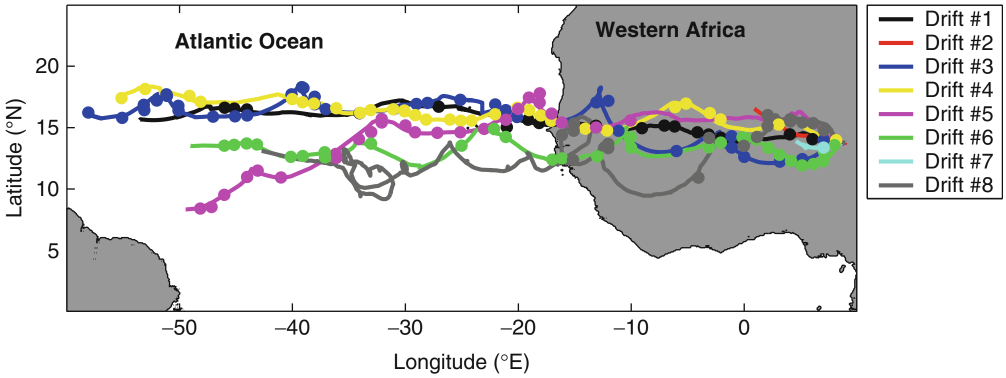
- Case identification (weather threat) using forecasts
- Computation of objective targeting guidance to help the decision process (where and when to deploy)
- Design of the deployment and decision making, transmitting the deployment instructions to the observation providers (observing platforms)
- Deployment of additional observations (data collection, transmission)
- Assimilation (this stage is automatic and do not imply any specific effort)
- A posteriori evaluation of the data impact (in research mode, optional, often ignored)

The Hurricane Synoptic Flow was the first regular seasonal TC targeting campaign, conducted by NOAA, on the North Atlantic hurricane basin (1982–1996). The research aircraft P3 deployed dropsondes along a circular navigation around the cyclone [5]. The choice of deployment was essentially subjective, but the campaign was a success. From 1997 to 2011, the NOAA Hurricane Synoptic Surveillance has combined subjectively and objectively guided deployments [1] of dropsondes in the Gulf of Mexico, Eastern Pacific, and Western Atlantic. In Asia, the annual campaigns DOTSTAR [25] that started in 2003 allow the improvement of the observing system in the Western Pacific with dropsondes released from Taiwanese jet aircrafts. Again, the choices of deployments are based on both subjective and objective analyses. From an NWP impact point of view, observations that are deployed in the close TCs' environment produce the best results in terms of TC trajectory prediction [11]. However, this conclusion holds for already formed and named TCs. Another issue is to better monitor tropical environment in which TCs may appear, develop, and strengthen.

Introduction

More recently, in West Africa, drifting balloons launched from Zinder (Niger) (Fig. 1) into the stratosphere (about 50 hPa) dropped 124 dropsondes over wide swaths of Africa and the Atlantic Ocean between end of August and end of September 2006, in order to document the meteorological environment of mesoscale convective systems (MCS) and TCs in the Atlantic [6]. This deployment of the so-called driftsonde system was performed in the frame of the African Monsoon Multidisciplinary Analysis (AMMA) program, initiated in 2002 with an intensive field campaign in 2006, and focusing on the physical changes in the environment of the West African monsoon and their impacts on society [20]. MCSs are the main precursors of TCs in the Atlantic (e.g., [2]), and may even contribute to tropical cyclogenesis in the Pacific (e.g., [10]). It is believed that scale interaction processes drive the tropical cyclogenesis that occurs over the Atlantic off the West African coast during late summer. This interaction involves:

- The African Easterly Jet (AEJ) develops at about 600-hPa pressure level and 4°N-5°N latitude because heating of the West African land mass during summer creates a surface temperature and moisture gradient between the Gulf of Guinea and the Sahara, and the atmosphere responds by generating vertical wind shear to maintain thermal wind balance.
- The African easterly waves (AEWs) which are synoptic scale, westward propagating disturbances produced by barotropic and baroclinic instabilities embedded in the AEJ.
- The Saharan air layer (SAL).
- The westward propagating MCSs.



Hurricane and Monsoon Tracking with Driftsondes. Figure 1
Trajectories of the eight driftsondes launched from Zinder (Niger) between August 28 and September 22, 2006. The *dots* indicate the location of a dropsonde

It is already known that some Atlantic tropical cyclones have their origin in MCSs propagating over the African continent [23]. For example, Hill and Lin [13] related the genesis of Hurricane Alberto (2000) to a MCS initiated over the Ethiopian highlands. Lin et al. [17] documented this case study and identified three successive convective genesis and lysis periods before the final cyclogenesis event off the Guinea coast. Berry and Thorncroft [3] also showed that the MCSs associated with the pre-Alberto disturbance were embedded in an AEW. The SAL may significantly impede the development of incipient tropical waves or disturbances and can even weaken preexisting tropical cyclones.

During the AMMA, the driftsondes sampled squall lines in the troughs of AEW and the southeastern edge of the tropical storm Florence, hurricanes Gordon and Helene. In addition to tracking potential hurricanes, the driftsondes gathered bird's-eye data on surges of hot, dry air that cascade into the Atlantic from the Sahel region of Africa. These surges carry huge amounts of dust as far west as Florida, influencing air chemistry, upper-ocean biology, and Atlantic weather systems. Although driftsondes have been tested over the last few years, this was the first time they were used in weather research and prediction. Based on this new type of dataset, this entry aims at illustrating and analyzing the potential of such observation system for tracking high-impact weather system such as cyclone and hurricanes in area generally void of in situ data, and provides additional observation for real-time forecast improvement.

Section"The Driftsonde Observations" presents the driftsonde system which was first used in the frame of a large scientific research program. Section"Analysis of the Driftsonde Trajectories and the Respective Synoptic Environment" describes the synoptic environment (including AEW activity) of the meteorological processes (MCS and tropical storm initiation) investigated with the series of vertical profiles provided by the dropsondes. Section "Analyses of AEW with the Driftsondes" describes the AEW activity and spatial pattern in relation with the dropsonde location. Section"Conclusion" concludes the description of the driftsonde system, lists the ongoing research activity using the driftsonde data collected during AMMA, and proposes perspectives for the use of such observation system.

The Driftsonde Observations

The concept of using driftsondes to take measurements over remote but scientifically important locations around the globe comes from THORPEX, a global program of the World Meteorological Organization (WMO) to accelerate improvements in the prediction of high-impact weather. A driftsonde is composed of a drifting stratospheric balloon and a gondola carrying dropsondes.

For the AMMA project, the drifting stratospheric balloons were developed by the Centre National d'Études Spatiales (CNES) whereas the dropsondes were designed at the National Center for Atmospheric Research (NCAR) and their development was funded by the National Science Foundation (NSF) and the National Oceanic and Atmospheric Administration (NOAA). To build the driftsonde system, scientists and engineers at NCAR and CNES had to overcome many hurdles. Each driftsonde had to be robust enough to endure days of extreme stratospheric cold (averaging −60°C) as well as the intense sunlight of the high, thin atmosphere. For the balloon deployment to be affordable and practical, the system also required low-cost, lightweight, off-the-shelf instruments capable of operating reliably in low pressure and in temperature extremes with very low power. Each gondola held about 35 dropsondes designed by NCAR (the number of dropsondes could differ between the driftsondes) carried on the ballooning systems designed by CNES. In order to make the driftsonde concept practical, NCAR developed a highly compact instrument package, roughly the size of a small bottle of water but weighing only about 140 g. Called MIST (Miniature In situ Sounding Technology), it weighs less than half as much as older dropsondes, which were designed at NCAR in the 1990s. Although driftsondes have been tested over the last few years, this was the first time they were used in weather research and prediction.

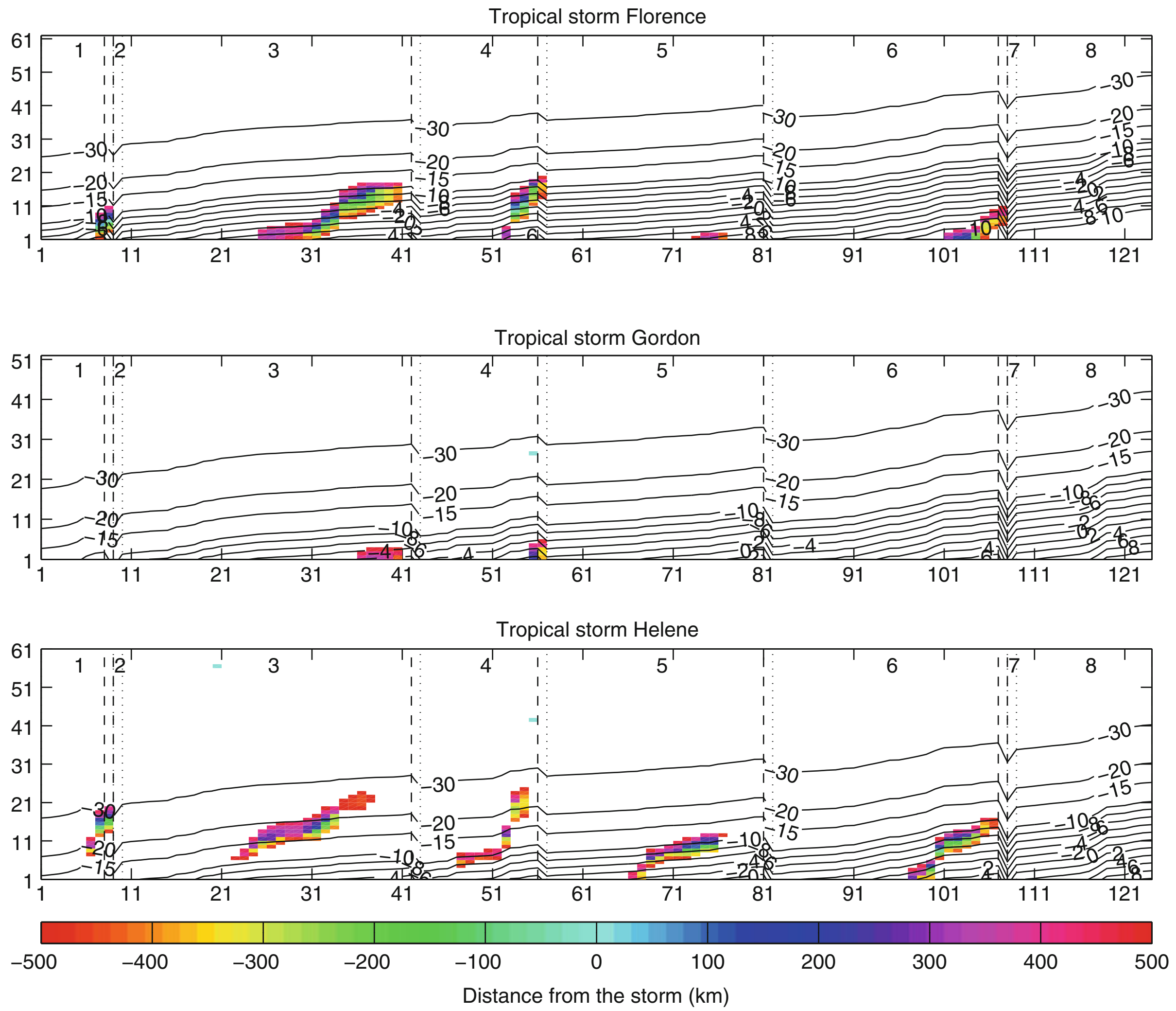
The Zinder site, located 740 km east of Niamey, was selected in order to study the AEW, that serve as seedlings for hurricanes during the late African monsoon period (August–September). Dozens of these waves moved across Africa into the Atlantic between about 10°N and 20°N. A small number developed into tropical storms. During July–September 2006 period, a total of 27 AEWs were objectively analyzed (compared with 31 in 2004 and 28 in 2005). July was very different from the later months with six out of seven of the first waves forming close to the longitude of Niamey or west of it. In August, the AEWs were initiated further east, between 10°E and 20°E, but AEWs over Niamey were still weak. Starting at the end of August and going into September, the AEWs became more coherent with stronger amplitudes over most of tropical North Africa. Interestingly, several AEWs also appeared to start further east, between 20°E and 30°E at this time. It should also be noted that all seven of the AEWs that became named tropical cyclones (Chris – AEW number 6; Ernesto – AEW number 13; Debby – AEW number 14; Florence – AEW number 18; Gordon – AEW number 19; Helene – AEW number 20; Isaac – AEW number 23; see [14]) were initiated east of Niamey and six of these occurred after the middle of August. After being launched from Zinder, each balloon drifted from Africa toward the Caribbean at heights of around 20 km height, where light easterly winds prevail. The trajectories exhibited cycloid-like patterns due to the presence of near-inertial waves [12] (Fig. 1). At least twice per day (0000 and 1200 UTC), each gondola released a dropsonde that fell by parachute, sensing the weather conditions during its 20-min descent and radioing data back to the gondola and then, by satellite, to the operation center in Paris. Whenever promising weather system developed, the operation center signaled the gondola to release additional dropsondes as often as in 3-h intervals. Eight driftsondes were released from Zinder during the late African monsoon period coinciding with the peak period for hurricane formation over the tropical Atlantic (August–September) and 124 sondes were successfully dropped from the eight driftsondes with 15 vertical profiles from the last dropsondes of driftsonde 8 sent to the GTS for assimilation. However, the number of successfully dropped sondes differed between driftsondes. Table 1 summarizes the operation during the driftsonde deployment. Table 1 shows that the most successful flights corresponded to driftsondes 1, 3, 4, 5, 6, and 8. However, the synoptic environment documented with driftsonde 1 was not of high scientific interest. Driftsonde 8 trajectory was particularly complex and difficult to manage since the flight period corresponded to the weakening of the 50 hPa easterly winds due to a change of the Madden Julian oscillation (MJO) phase. So driftsonde 8 did not allow the tracking of any interesting meteorological event. In the following, we thus analyze in detail the data collected with driftsondes 3–6. They documented in an unprecedented way AEW initiating continental mesoscale convective systems (MCS) evolving over the ocean into tropical storm like Florence and hurricanes like Gordon and Helene, some of them skirting with the US Atlantic and Gulf coasts and all experiencing extra-tropical transition.

Analysis of the Driftsonde Trajectories and the Respective Synoptic Environment

Figure 2 shows the space-time “distances” of the dropsondes from the center of tropical storms

Hurricane and Monsoon Tracking with Driftsondes. Table 1 Summary of driftsonde operations

Driftsonde number	First dropsonde coordinates and date	Last dropsonde coordinates and date	Number of successful dropsondes
1	(6.04°E,14.05°N) – August 28 1406 UTC	(−44.03°E,16.47°N) – September 2 1217 UTC	8
2	(8.13°E,13.93°N) – August 29 2118 UTC	(8.13°E,13.93°N) – Augember 29 2118 UTC	1
3	(7.04°E,13.99°N) – September 1 2351 UTC	(−58.07°E,16.22°N) – September 9 1312 UTC	33
4	(8.14°E,13.97°N) – September 4 1942 UTC	(−55.14°E,17.40°N) – September 11 1110 UTC	14
5	(−10.07°E,15.58°N) – September 9 0807 UTC	(−50.09°E, 8.39°N) – September 13 1808 UTC	25
6	(8.00°E,13.57°N) – September 9 2055 UTC	(−46.03°E,13.52°N) – September 18 0005 UTC	26
7	(7.03°E,13.37°N) – September 12 2358 UTC	(7.03°E,13.37°N) – September 12 2358 UTC	1
8	(6.10°E,15.36°N) – September 16 0603 UTC	(−15.07°E,15.33°N) – September 22 1742 UTC	16


Tropical storm Florence

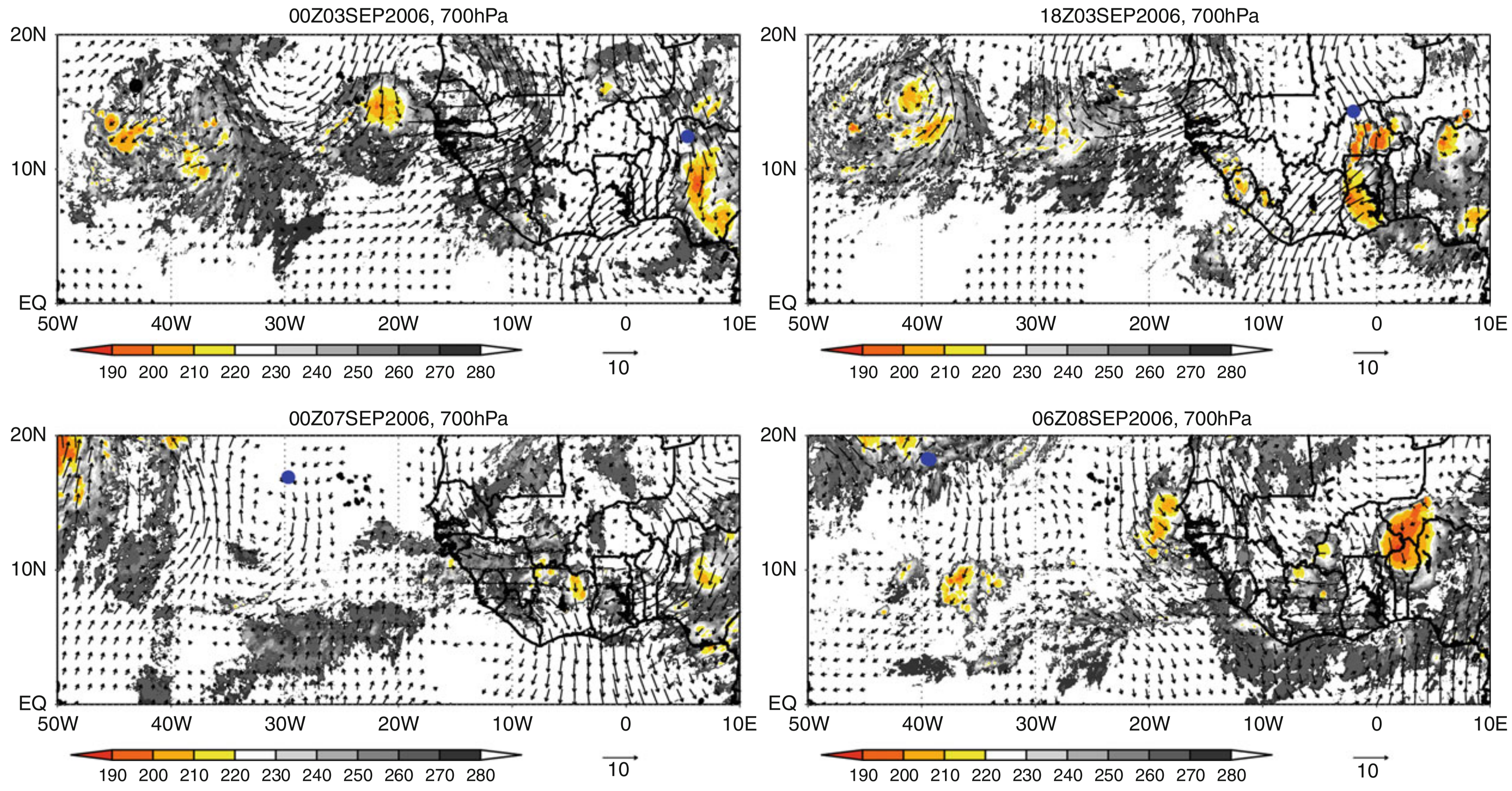
Tropical storm Gordon

Tropical storm Helene

Hurricane and Monsoon Tracking with Driftsondes. Figure 2
Space (color code) and time (iso-contours) "distances" of the dropsondes from the center of tropical storms (TS) Florence (**a**), Gordon (**b**), and Helene (**c**) (color code). The abscissa corresponds to the number of space-time coordinate information available for the TS trajectory. The ordinate corresponds to the number of successful dropsondes available for the whole field campaign. The numbers at the top of the figure indicate the identification number of the driftsonde. Only distances smaller than 500 km are shown, the positive (negative) sign indicates that the sonde is dropped ahead (behind) of the center of the storm. The iso-contours are shown for −30, −20, −15, −10, −8, −6, −4, −2, 0, 2, 4, 6, 8, and 10 days, the positive (negative) sign indicate that the sonde is dropped before (after) the storm

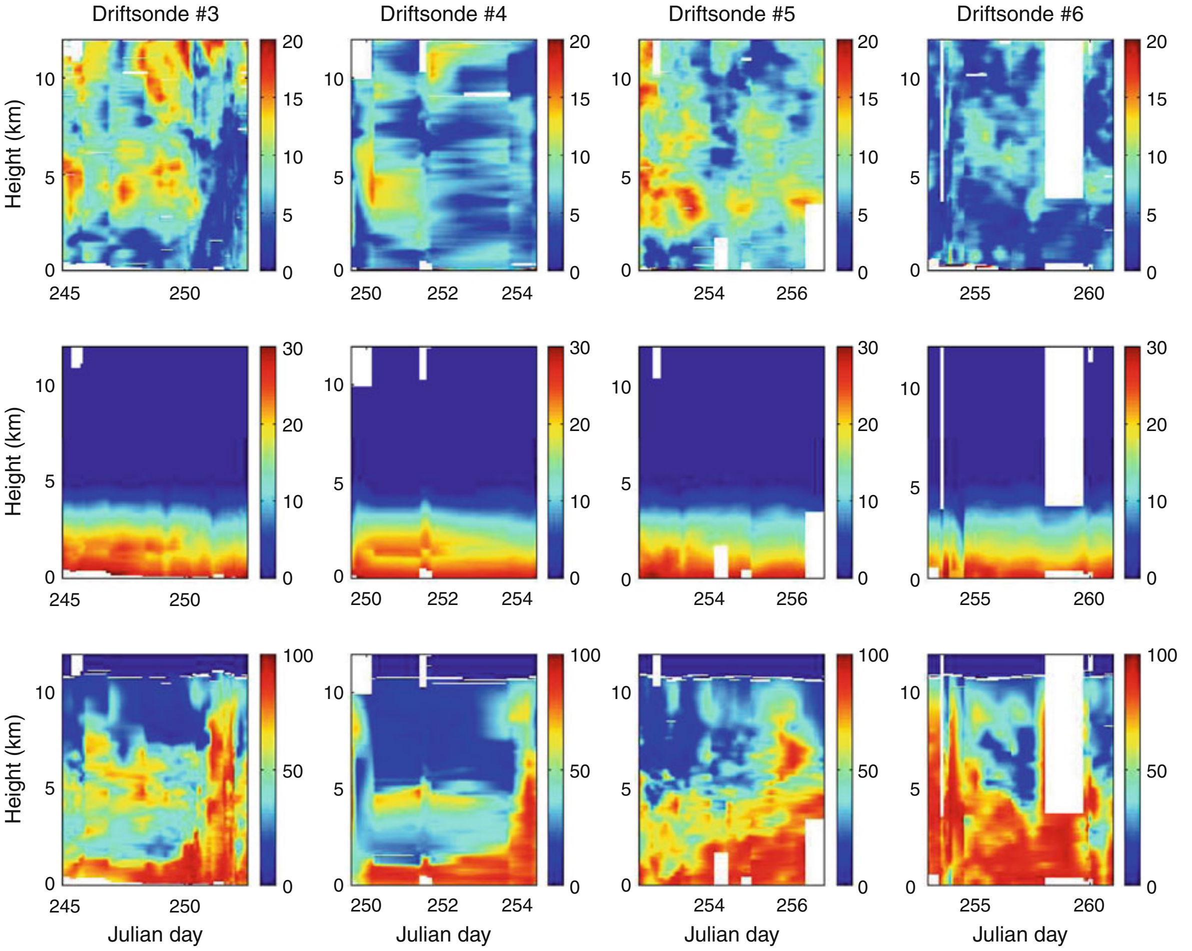
Florence, Gordon, and Helene. The closest in space and time from the storm center is the best. Figure 2a shows that driftsonde 1 approaches Helene track at less than 500 km about a week in advance, whereas driftsondes 3 and 4 followed tropical storm Florence dropping several sondes less than 500 km from the storm track. Gordon track was sampled by driftsonde 4 1 day after its passage (Fig. 2b) whereas Helene track was sampled by driftsonde 6 less than 1 day after its passage (Fig. 2c).

Figure 3 displays the Meteosat Second Generation (MSG) 10.8 μm channel brightness temperature which indicates that convective activity as well as the filtered 700 hPa wind field is shown with arrows (the absence of arrows indicate winds weaker than 1 m s^{-1}). It shows



Hurricane and Monsoon Tracking with Driftsondes. Figure 3

Meteosat Second Generation (MSG) 10.8 μm channel brightness temperature which indicates convective activity with superimposed filtered 700 hPa wind field shown with *arrows* (the absence of arrows indicate winds weaker than 1 m s^{-1}) on September 3, 2006, (Julian day 246) at 0000 UTC (**a**), on September 3, 2006, (Julian day 246) at 1800 UTC (**b**), on September 7, 2006, (Julian day 250) at 0000 UTC (**c**) and on September 8, 2006, (Julian day 251) at 0600 UTC (**d**). The black and blue filled dots indicate the locations of dropsondes from driftsondes 1 and 3, respectively that driftsonde 3 drifts in the vicinity of an MCS propagating westward over the continent at about $15\ \mathrm{m\ s^{-1}}$. At about 40°W, tropical depression Florence is forming. Figure 4 (first column) shows a time versus height cross section of wind speed, temperature, and humidity constructed with the vertical profiles obtained by the sondes dropped from driftsonde 3. Figure 4a shows evidence of a strong wind tongue between 4 and 5 km height corresponding to the AEJ. Figure 4i shows the moist African planetary boundary layer between Julian days 245 and 248 (September 2–5, 2006) (75% relative humidity) extending up to about 1 km height. Above the planetary boundary layer, layers of saturated air between 5 and 8 km height reveal the proximity of the convective activity. On Julian day 250 (September 7, 2006), dropsondes are released from driftsonde 3 few hundreds of kilometers downstream of Florence, located at about 50°W, which in the meantime has been classified as a tropical storm (it becomes a hurricane on September 10, 2006). The near-surface wind decreases in the disturbed environment near tropical storm Florence as shown in Fig. 4a while the planetary boundary layer humidity and depth increase (up to about 90% relative humidity and up to 3 km



Hurricane and Monsoon Tracking with Driftsondes. Figure 4
Time versus height cross section of wind speed (*first row*, panels **a**, **b**, **c**, and **d**), temperature (*second row*, panels **e**, **f**, **g**, and **h**) and humidity (*third row*, panels **i**, **j**, **k**, and **l**) constructed with the vertical profiles obtained by the sondes dropped from driftsonde 3 (first column; blue trajectory in Fig. 1), 4 (second column; yellow trajectory in Fig. 1), 5 (third column; purple trajectory in Fig. 1), and 6 (fourth column; green trajectory in Fig. 1)

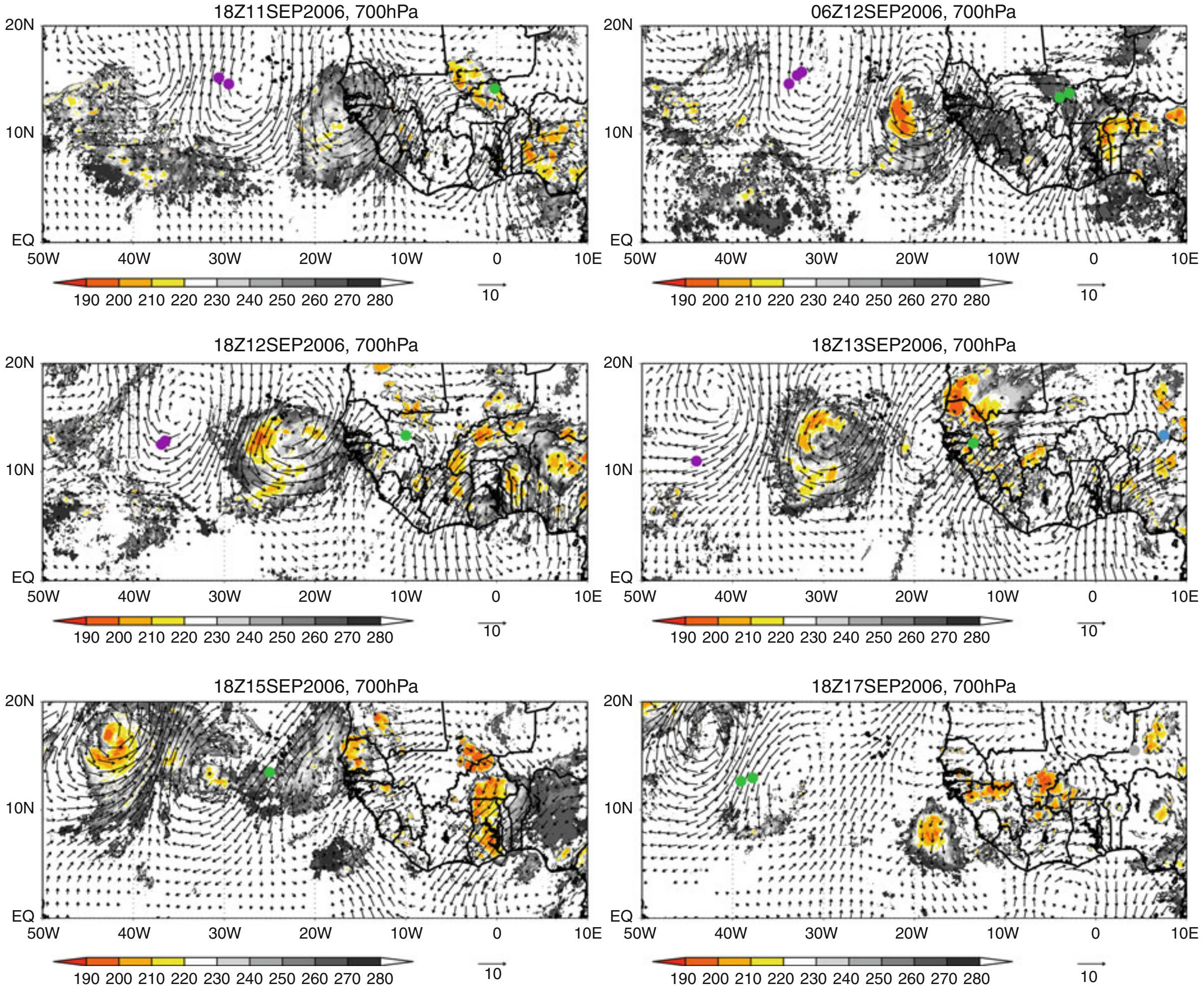
during day 250) before reaching the disturbed environment near Florence where very intense convection occurs (Fig. 4i). Figure 4b is similar to Fig. 4a for driftsonde 4. However, the much lower number of successful dropsondes does not allow the documentation of fine-scale structures over the Atlantic Ocean. It however shows a disturbed environment similar to driftsonde 3 less than 200 km south of the developing tropical storm Gordon at about 55°W (not shown) and Julian day 254 with large humidity extending up to the tropopause and weak near-surface winds. The existence of weak-surface winds is surprising since one may associate naturally developing tropical storms with large evaporation and thus strong winds. The vertical profiles of relative humidity documented by the dropsondes of driftsonde 4 around Julian day 250 also show evidence of very low humidity values just above the planetary boundary layer (around 20% relative humidity) and at about 5–6 km height (below 10% relative humidity) corresponding the AEJ (Fig. 4). The low level dry layer was identified as a strong dry air outflow from the Sahara by the Meteosat-8/Goes-10 combined Saharan air layer product. The very dry air conveyed by the AEJ originated in the upper levels (200–250 hPa) on the anticyclonic side of the polar jet stream at 50°N as diagnosed by the method proposed by Roca et al. [21].

Figure 5 is similar to Fig. 3 between September 11, 2006, and September 17, 2006. Driftsondes 5 and 6 represented by purple and green filled dots, respectively, sensing the atmosphere ahead and downstream the tropical storm Helene which initiates on September 11, 2006, at 1800 UTC over the coast of Senegal and moves westward at a propagation speed of about 8 m s^{-1}. Figure 4 shows that after the formation of three successive storms in about 2 weeks, all evolving into hurricane category, driftsondes 5 and 6 probe a much more disturbed environment than for Florence, with higher convective activity both over the continent and the ocean with nearly saturated air observed up to 5 km along the driftsonde tracks. The AEJ tends to weaken during driftsonde 6 flight.

Analyses of AEW with the Driftsondes

As AEWs are the dominant synoptic weather systems in West Africa and the tropical Atlantic during boreal summer and are an important component of the regional climate by modulating West African rainfall including mesoscale convective systems (MCSs) which can be precursors of tropical cyclones in the Atlantic, the respective locations of the dropsonde and AEW are now investigated.

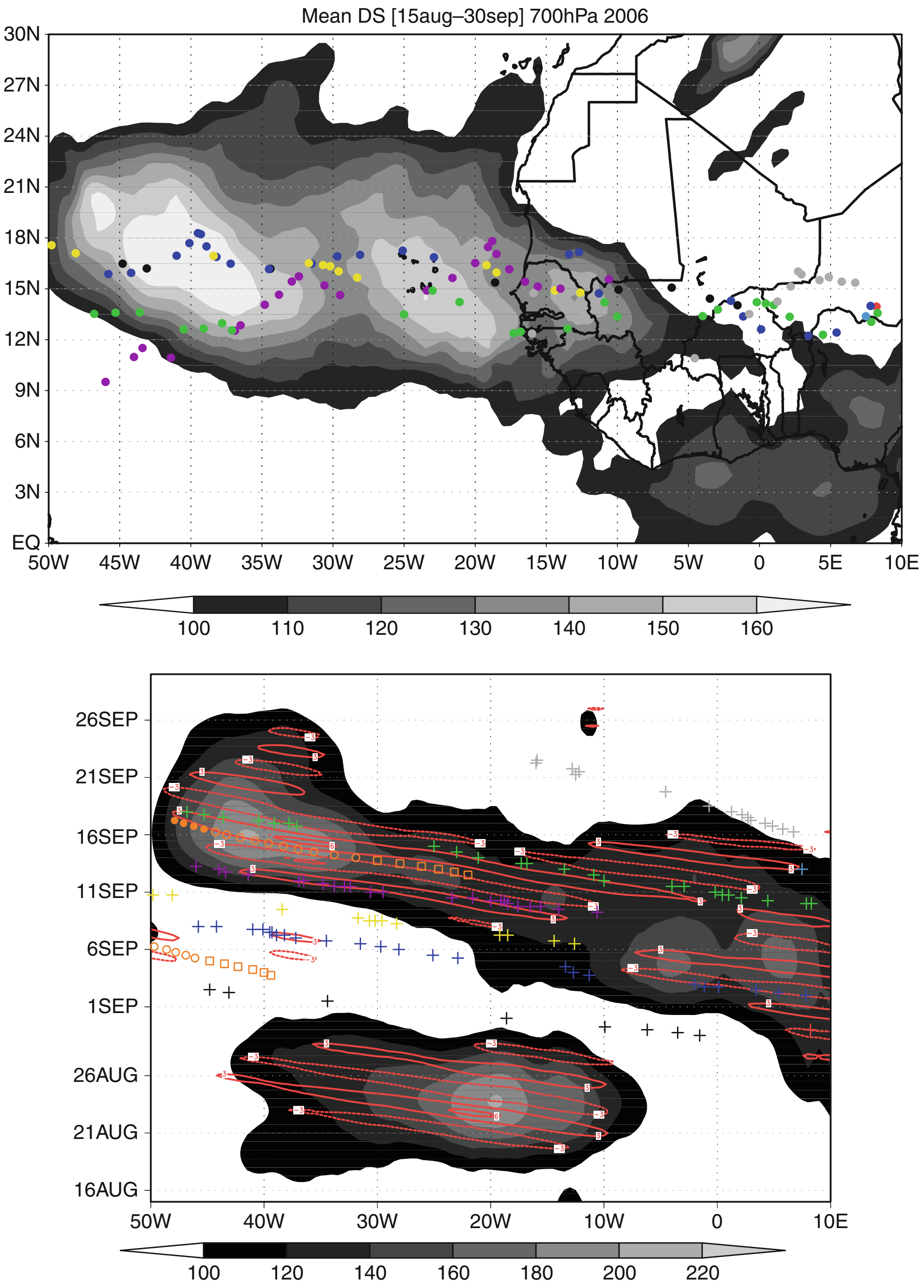
A method based on the wavelet analysis with a Morlet mother wavelet [24] of the meridional wind field at 700 hPa is used to detect AEW activity. The sum of the spectral density between 3 and 5 days period allows to perform an index which estimates the wave activity over each grid points. The spatial average of this index is then computed to observe the large-scale AEWs activity. This method has been proposed by Lavaysse et al. [16]. Figure 6 shows the longitude/latitude field of the mean AEW index averaged between August 15 and September 30, 2006, as well as the Hovmöller space-time diagram of the 2–6 day bandpass filtered meridional wind at 700 hPa averaged between 5°N and 20°N. Using this method, we detect in 2006 three periods of large AEW activity over West Africa and the tropical Atlantic (i.e., 50°W, 10°E and 5°N, 20°N): middle of July, middle of August, and middle of September. This is in accordance with other methods of AEW as the variance of the 700 hPa meridional wind filtered in the 3–5 days band period, and the 700 hPa curvature vorticity method by Berry et al. [4] and illustrated by Janicot et al. [14]. These results are also in accordance following different analyses as ECMWF and national centers for environmental predictions (NCEP) (not shown) and different geopotential levels (700 or 850 hPa). During the period of driftsonde operations (i.e., August 29 to September 23, 2006), the AEW activity shows two distinct periods. Driftsonde 1 (black trajectory, Fig. 1) flies during a period of low AEW activity preceding the initiation of Helene tropical storm on September 3, 2006, at 40°W also during the period of low AEW activity. The AEW activity increases on September 1, 2006, over West Africa. The area of large waves activity propagates westward in association with the westward propagation of the filtered meridional wind. Driftsonde 3 (blue trajectory) follows the ridge sector of the strengthening wave train (Fig. 3) but the AEW activity abruptly decreases when the waves reach the ocean (around 20°W). A similar scenario occurs for driftsonde 4 (yellow trajectory). Around September 7, 2006, AEW



Hurricane and Monsoon Tracking with Driftsondes. Figure 5

Meteosat Second Generation (MSG) 10.8 μm channel brightness temperature which indicates convective activity with superimposed filtered 700 hPa wind field shown with *arrows* (the absence of *arrows* indicate winds weaker than 1 m s^{-1}) on September, 11, 2006, (Julian day 254) at 1800 UTC (**a**), on September 12, 2006, (Julian day 255) at 0600 UTC (**b**), on September 12, 2006, (Julian day 255) at 1800 UTC (**c**), on September 13, 2006, (Julian day 256) at 1800 UTC (**d**), on September 15, 2006, (Julian day 258) at 1800 UTC, (**e**) and on September 17, 2006, (Julian day 260) at 1800 UTC (**f**). The *purple*, *green*, and *cyan filled dots* indicate the locations of dropsondes from driftsondes 5, 6, and 7, respectively

activity is maximum over West Africa. In the same time, driftsonde 5 (purple trajectory) is launched (Table 1). Ten sondes are dropped in the ridge sector of the AEW over West Africa (Fig. 5). Above the coast, the westward propagation speeds of the driftsonde and the AEW are different thus on September 11, driftsonde 5 drops sondes in the trough sector of the AEW ahead of the ridge sector. Driftsonde 5 is particularly interesting when flying over the continent because of its proximity to the trough of the AEW associated with large convective activity. This AEW-MCS system is the precursor of tropical storm Florence initiating around September 12, at 20°W. This storm propagates along a maximum of the mean AEW activity index with the largest value over the tropical Atlantic (September 15 at 42°W; Fig. 6). Driftsonde 6 (green trajectory) covers the largest longitudinal cross section with a first drop at 7°E and the last drop at 47°W. Driftsonde 6 remains



Hurricane and Monsoon Tracking with Driftsondes. Figure 6

Left panel: Mean AEW index with the location of the dropsondes superimposed. *Right panel*: Hovmöller space-time diagram of the 2–6 day bandpass filtered NCEP-2 meridional wind at 700 hPa (*red contour* for +1 m s^{-1}) and AEW index (*gray shading*) averaged between 5°N and 20°N. Superimposed are the locations of the dropsondes of the eight driftsondes ("+" marker with same color code as in Fig. 1) and the 1-day averaged location of the tropical storms Florence and Gordon (*squares*, *circles*, and *filled circles* when classified as tropical depression, tropical storm, and hurricane, respectively). Gordon is not visible since forming at about 55°W, that is, out of the abscissa scale over the same ridge sector along the whole operation period and provides temporal-longitudinal evolution of the activity and characterization of the ridge.

Conclusion

It is worth emphasizing that all the driftsondes deployed during AMMA were still on probation. This status explains the late dissemination of the data on the GTS (driftsonde 8, only) and the technically mitigated success. The driftsonde system has gradually reached the stage of maturity. During T-PARC, 16 driftsondes were launched from Hawaii during August 2008 [9]. About 250 dropsondes were targeted eventually that correspond to a 64% success rate. The launch site of the pre-Concordiasi test-campaign was in the Seychelles in February 2010 and 3 driftsondes with limited dropsondes were deployed. Finally, Concordiasi test-campaign took place from September to December 2010 [19] over Antarctica. More than 620 dropsondes were deployed from a fleet of 13 driftsondes launched at McMurdo station. The mission success rate was close to 95%, most of the data were disseminated on the GTS.

Whenever the dropsonde data are disseminated or not, the evaluation of the NWP-wise benefit of deploying these observing platforms requires some specific impact studies. These imply to compare the quality of forecasts when the NWP system processes the dropsondes' data or not. The data collected over West Africa and the Atlantic Ocean during AMMA have been used to evaluate numerical weather prediction of the late 2006 African monsoon [8]. They will also be used for a posteriori analysis of their added value for numerical weather forecast. The evaluation of NWP impact of T-PARC driftsondes has not been carried out yet, though many phenomenological studies have been carried out. The evaluation of Concordiasi is currently being done. As the NWP benefit is gained as soon as the data are assimilated in operational NWP systems, the exact assessment of this benefit triggers a limited motivation. Indeed, this evaluation is crucial to correctly assess the NWP efficiency of data targeting with such platforms.

Future Directions

The idea of a network of high atmosphere balloons distributed in the stratosphere and dropping sondes on demand has existed for more than a decade. Some projects have emerged with the advent of THORPEX. Many of them were intractable. The driftsonde system, which appeared to be one of the less ambitious projects, eventually emerged in AMMA 2006. The system has been improved in T-PARC 2008 and for Concordiasi in 2010. The system will be upgraded in 2011–2013, but this last technical step will prevent the system to be deployed in the joint T-NAWDEX/HyMeX framework in 2012–2014 (T-NAWDEX stands for THORPEX-North Atlantic Waveguide and Downstream impact Experiment and HyMeX means Hydrological Cycle in the Mediterranean Experiment). In this last field campaign, the challenge would have been to deploy driftsondes above the Atlantic Ocean, from the USA. The objective was to sample the whole Atlantic, upstream from Europe and the Mediterranean region. Hopefully, the new generation system will probably be deployed in further future specific observation campaigns.

In near future and NWP-wise, the main interest of the driftsonde system is its ability to sample a whole atmospheric environment in which a weather threat is likely to emerge. This mode is especially effective if driftsondes are deployed in a network of several units. This is typically the level of targeting that should be achieved for adaptive observation. Indeed, several factors contribute to the suboptimality of single-use (single weather event) driftsondes' deployment.

- The release of driftsonde is a delicate art, highly dependent on environmental conditions. The system has a limited capacity to respond to a very specific request (short time and very localized area).
- After having drift above the area of interest as defined for the targeted event, the balloon is lost to its initial targeting mission.
- Controlling the trajectory is currently not possible. In the cases and T-PARC and AMMA (summer) the trajectories showed rather zonal despite a few fluctuations. But during the probe campaign for Concordiasi in the Seychelles (February 2010), the trajectories were poorly predictable and let the balloons drift out of the Tropics.

However, regular sampling of regions void of any observations can be very profitable for weather forecasting. Furthermore, deployment in areas (space-time) where the flow is periodic preserves driftsondes' fleet on the region of interest: around the globe inter-tropical circulation in summer or polar vortex in winter. The specific impact remains to be shown clearly with the existing datasets. This should be a prerequisite to any further deployment and the time left by the upgrade of the driftsonde system should be used to demonstrate the usefulness of these data in some global prediction systems.

Finally, there are other aerostats that may be more suitable for adaptive deployment in response to the possible occurrence of some weather event in particular. Boundary layer pressurized balloon (BLPB) is one example. The BLPB is a balloon drifting in the atmospheric boundary layer and carrying an instrumented gondola which collects measurements of the typical meteorological variables. The proof of concept will be given during the HyMeX special observing periods. In the HyMeX field campaigns, the BLPBs will be deployed above the marine surroundings of some intense convective systems that occur at the periphery of the Mediterranean basin [7].

Acknowledgments

We are thankful to J. Fox, K. Romberg, J. VanAndel, H. Cole, C. Martin, G. Granger, D. Flanigan, and D. Lauritsen (NCAR) for assistance in operation of the driftsonde system. Based on French initiative, AMMA was built by an international scientific group and is currently funded by a large number of agencies, especially from France, UK, Germany, USA, and Africa. It has been the beneficiary of a major financial contribution from the European Community's Sixth Framework Research Program. Detailed information on scientific coordination and funding is available on the AMMA International website http://www.amma-international.org.

Bibliography

1. Aberson SD (2003) Targeted observations to improve operational tropical cyclone track forecast guidance. Mon Wea Rev 131:1613–1628
2. Avila LA, Pasch RJ (1992) Atlantic tropical systems of 1991. Mon Wea Rev 120:2688–2696
3. Berry GJ, Thorncroft C (2005) Case study of an intense African easterly wave. Mon Wea Rev 133:752–766
4. Berry GJ, Thorncroft C, Hewson T (2007) African easterly waves during 2004 – analysis using objective techniques. Mon Wea Rev 135:1251–1267
5. Burpee RW, Franklin JL, Lord SJ, Tuleya RE, Aberson SD (1996) The impact of omega dropwindsondes on operational hurricane track forecast models. Bull Am Meteorol Soc 77:925–933
6. Drobinski P, Parsons D, Cocquerez P, Cardonne A, Cole H, Fox J, Lefevre JP, Redelsperger JL, Romberg K, VanAndel J, Vénel S (2006) Des ballons stratosphriques traquent la mousson africaine. La Météorologie 55:2–3
7. Drobinski P, Basdevant C, Doerenbecher A, Pannekoucke O, Fesquet C, Verdier N, Vargas A (2010) Balloon deployment during the HyMeX (Hydrological in the Mediterranan EXperiment) project. 38th COSPAR Scientific Assembly, Bremen
8. Drobinski P, Karbou F, Bauer P, Cocquerez P, Lavaysse C, Hock T, Parsons D, Rabier F, Redelsperger JL, Vénel S (2011) Driftsonde observations to evaluate numerical weather prediction of the late 2006 African monsoon. J Appl Meteorol Climatol (submitted)
9. Elsberry RL, Harr PA (2008) Tropical cyclone structure (TCS-08) field experiment science basis, observational platforms, and strategy. Asia-Pacific J Atmos Sci 44:209–231
10. Frank NL (1970) Atlantic tropical systems of 1969. Mon Wea Rev 98:307–314
11. Harnisch F, Weissmann M (2010) Sensitivity of typhoon forecasts to different subsets of targeted dropsonde observations. Mon Wea Rev 138:2664–2680
12. Hertzog A, Vial F, Mechoso CR, Basdevant C, Cocquerez P (2002) Quasi-Lagrangian measurements in the lower stratosphere reveal an energy peak associated with near-inertial waves. Geophys Res Lett 29:1229. doi:10.1029/2001GL014083
13. Hill CM, Lin YL (2003) Initiation of a mesoscale convective complex over the Ethiopian Highlands preceding the genesis of Hurricane Alberto (2000). Geophys Res Lett 30:1232. doi:10.1029/2002GL016655
14. Janicot S, Thorncroft CD, Ali A, Asencio N, Berry G, Bock O, Bourles B, Caniaux G, Chauvin F, Deme A, Kergoat L, Lafore JP, Lavaysse C, Lebel T, Marticorena B, Mounier F, Nedelec P, Redelsperger JL, Ravegnani F, Reeves CE, Roca R, de Rosnay P, Schlager H, Sultan B, Tomasini M, Ulanovsky A (2008) ACMAD forecasters team: large-scale overview of the summer monsoon over West Africa during the AMMA field experiment in 2006. Ann Geophys 26:2569–2595
15. Joly A, Browning KA, Bessemoulin P, Cammas J-P, Caniaux G, Chalon J-P, Clough SA, Dirks R, Emanuel KA, Eymard L, Gall R, Hewson TD, Hildebrand PH, Jorgensen D, Lalaurette F, Langland RH, Lematre Y, Mascart P, Moore JA, Persson PO, Roux F, Shapiro MA, Snyder C, Toth Z, Wakimoto RM (1999) Overview of the field phase of the Fronts and Atlantic Storm-Track Experiment (FASTEX) project. Bull Am Meteorol Soc 125:3131–3163
16. Lavaysse C, Diedhiou A, Laurent H, Lebel T (2006) African easterly waves and convective activity in wet and dry sequences of the West African Monsoon. Clim Dyn 27: 319–332
17. Lin YL, Robertson KE, Hill CM (2005) Origin and propagation of a disturbance associated with an African easterly wave as

a precursor of Hurricane Alberto (2000). Mon Wea Rev 133:3276–3298

18. McNally T (2002) A note on the occurrence of cloud in meteorologically sensitive areas and the implications for advanced infrared sounders. Q J Roy Meteorol Soc 128:2551–2556
19. Rabier F, Bouchard A, Brun E, Doerenbecher A, Guedj S, Guidard V, Karbou F, Peuch V-H, El Amraoui L, Puech D, Genthon C, Picard G, Town M, Hertzog A, Vial F, Cocquerez P, Cohn SA, Hock T, Fox J, Cole H, Parsons D, Powers J, Romberg K, VanAndel J, Deshler T, Mercer J, Haase JS, Avallone L, Kalnajs L, Mechoso CR, Tangborn A, Pellegrini A, Frenot Y, Thpaut J-N, McNally A, Balsamo G, Steinle P (2010) The Concordiasi project in Antarctic. Bull Am Meteorol Soc 91:69–86
20. Redelsperger JL, Thorncroft CD, Diedhiou A, Lebel T, Parker DJ, Polcher J (2006) African monsoon multidisciplinary analysis: an international research project and field campaign. Bull Am Meteorol Soc 87:1739–1746
21. Roca R, Lafore JP, Piriou C, Redlesperger JL (2005) Extratropical dry-air intrusions into the West African monsoon midtroposphere: an important factor for the convective activity over the Sahel. J Atmos Sci 62:390–407
22. Szunyogh I, Toth Z, Zimin A, Majumdar SJ, Persson A (2002) On the propagation of the effect of targeted observations: the 2000 Winter Storm Reconnaissance Program. Mon Wea Rev 130:1144–1165
23. Thorncroft C, Hodges K (2001) African easterly wave variability and its relationship to Atantic tropical cyclone activity. J Clim 14:116–1179
24. Torrence C, Compo GP (1998) A practical guide to wavelet analysis. Bull Am Meteorol Soc 79:61–78
25. Wu C-C, Chen J-H, Lin P-H, Chou K-H (2007) Targeted observations of tropical cyclone movement based on the adjoint-derived sensitivity steering vector. J Atmos Sci 64:2611–2626

Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles

M. Abul Masrur[1], Vijay K. Garg[2]
[1]College of Engineering & Science, University of Detroit Mercy, Detroit, MI, USA
[2]Dublin, OH, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Off-Road Vehicles
Diesel Electric Locomotives
Hybrid Hydraulic Vehicles
Future Directions
Appendix
Bibliography

Glossary

Plug-in hybrid A plug-in hybrid is a regular hybrid vehicle that has a large high-capacity battery bank that can be charged by plugging in to normal household outlet. While standard hybrids require a combination of regenerative braking and energy from the engine to recharge the batteries, plug-ins can essentially operate as electric vehicles with an internal combustion engine backup. This is also known as PHEVs (Plug-in Hybrid Electric Vehicles)

Alternator An electromechanical device used to translate mechanical power into AC electrical power.

Battery state of charge SOC or state of charge is the equivalent of a fuel gauge for a battery, that is, a measure of how much energy is left in the battery with respect to its full rated capacity. Although many times SOC is related directly by measuring the terminal voltage, SOC cannot be accurately determined in general by voltage measurement, because the terminal voltage of a battery may stay substantially constant until it is completely discharged. SOC is also highly dependent on temperature.

Complex hybrid vehicles In this system, in addition to the ICE and the motor, there is an additional motor/generator set (along with a power electronic converter) which can be used to charge the battery as needed. This can continue charging the battery while the main motor is driving the wheels.

Converter A generic term for a system which translates one form of power (DC or AC) into another.

Diesel IC engine An IC engine that uses the heat of highly compressed mixture of air and spray of fuel (diesel) near the end of the compression stroke to create the combustion.

Drive cycle It is a relationship of either speed versus time or power versus time, for a particular vehicle which is driven in a given road profile. The road profile can be city streets, highways, steep grade, or some chosen mixed combination of these or other profiles, which will be predefined.

Fuel cell It is a stationary device used to directly translate chemical energy into electrical energy.

Generator It is the generic name used to imply something that translates mechanical energy into electrical energy.

Hybrid electric vehicle (HEV) A vehicle which uses more than one source (two being most common) of propulsion, one being the internal combustion engine, and the other uses an electric motor driven by either a generator run by the IC engine or a combination of generator and a battery which can be charged by the generator to provide the electricity, or it can be a fuel cell instead of a battery or generator, which can directly run the motor.

Internal combustion engine (ICE) ICE is the one in which combustion of the fuel (e.g., gasoline) takes place internally inside a cylinder or other enclosed space using spark ignition or compression ignition to some combustible gas, which will expand and produce high pressure that will produce mechanical power. ICE is thus a converter of chemical energy in mechanical energy.

Inverter A system which converts DC into AC.

Microgrid If an HEV is provided with an additional converter, which can translate the battery power into a utility level AC power, then it can be interfaced with either the main utility grid directly or with other HEVs or other smaller subsystems using utility level AC voltage. An interfaced set of vehicles, other utility level AC voltages, and utility grid, all or part of these entities taken together can be termed a microgrid.

Off-road vehicle An off-road vehicle is a vehicle, which is capable of driving on and off paved or gravel surface. Vehicles that do not travel public streets or highways are generally termed off-highway vehicles, for example, tractors, forklifts, cranes, backhoes, bulldozers, and golf carts.

Parallel hybrid vehicles In parallel hybrid, the IC engine and the electric motor, both are mechanically connected to the wheel. The ICE is also used to charge the batteries.

Power electronics It is a system using solid state high-power and high-voltage electronic switches to turn on/off a voltage or current, which together with other elements like inductance, capacitance to convert DC, AC, etc., into another form of controlled power.

Rectifier A system which converts AC into DC.

Series hybrid vehicles It is a hybrid vehicle, in which the electric motor (or motors) is the only unit which is used to run the wheels. The IC engine charges the batteries by means of a generator.

Ultracapacitor A capacitor that has much greater energy density and power per pound than regular capacitors. It uses chemicals, but the energy is stored in the electrostatic form and, unlike a battery, the energy is not stored through the chemical reaction and no change in chemicals take place while storing energy. It is also sometimes referred to as “supercapacitor.”

Definition of the Subject and Its Importance

Hybrid electric vehicle (HEV), as discussed elsewhere in this entry, basically involves hybridization or mixing of power and propulsion sources and/or loads, leading to the achievement of better overall system efficiency and accompanying pollution control. In short, it hinges on the principle that not every “type of” power source (which, e.g., could be IC engine, battery, generator, etc.) is always the best to provide power to every type of load demand, which can vary over time. So, the idea is to use the best source as the load changes. Here, the word “best” is defined in terms of efficiency, performance, and similar attributes. The technology evolved in different forms over the years. The main motivation behind the technology in the recent past has been due to cost of fossil fuel, that is, gasoline and diesel in particular, which showed an ever-increasing trend. But additional benefit like less pollution has also accompanied the technology. HEV technology has already been in existence for use in diesel electric locomotives for quite a while. There the intent was to get very high torque to move the vehicle from rest, by using electric motor instead of a mechanical system (with engine and transmission directly propelling the wheels), thus avoiding the use of bulky mechanical transmission system. At this time, the technology, which is highly interdisciplinary in nature, requires further improvement in the areas of battery, power electronics, and motor. To this end better materials are being investigated to develop lightweight and high temperature withstanding systems. The importance of the technology is obviously evaluated mainly by how much it can help with fuel economy, followed by pollution control, and then the performance, weight, and size. Depending on the application, one or more of these criteria can be of importance to one user (e.g., small-size automobiles), whereas as another set of criteria can be of importance to another user (e.g., heavy construction and off-road vehicles). Obviously, there is much to gain from the development of this important technology and the future seems very promising at this time with significant technical and social ramifications.

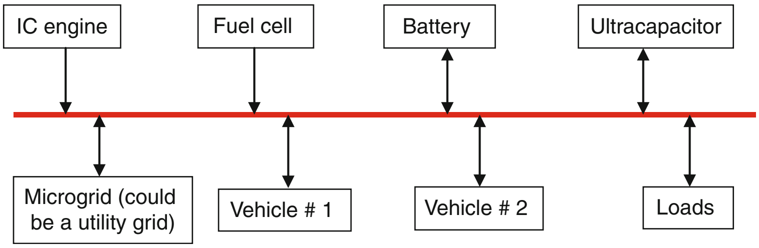
Introduction

There are many HEV (hybrid electric vehicle) applications which include both vehicular and electrically operated nonpropulsion-type military usages. Vehicle applications are directly related to vehicular propulsion. However, interfacing of the vehicular electrical system to create a microgrid to supply/receive power to/from utility grid is gaining popularity. Multiple HEVs, with proper interfacing mechanisms, can create utility-level voltages to run various stationary equipments. This can be done by running the HEV battery (and if necessary, the IC engines [ICE] in the HEV). In another application, several HEVs, along with an industrial/utility power system, can form a microgrid environment and help enhance the overall power system robustness and availability. Hybrid vehicles can be interfaced with a utility grid or a microgrid, where a number of vehicles could be supplying power through some common bus. It is, in principle, possible to have an IC engine–based generator, and fuel cell, vehicles, all properly synchronized and then exchanging power. A system-level diagram of such a situation is shown in Fig. 1. However, in this entry only off-road vehicles/machines applications will be discussed.

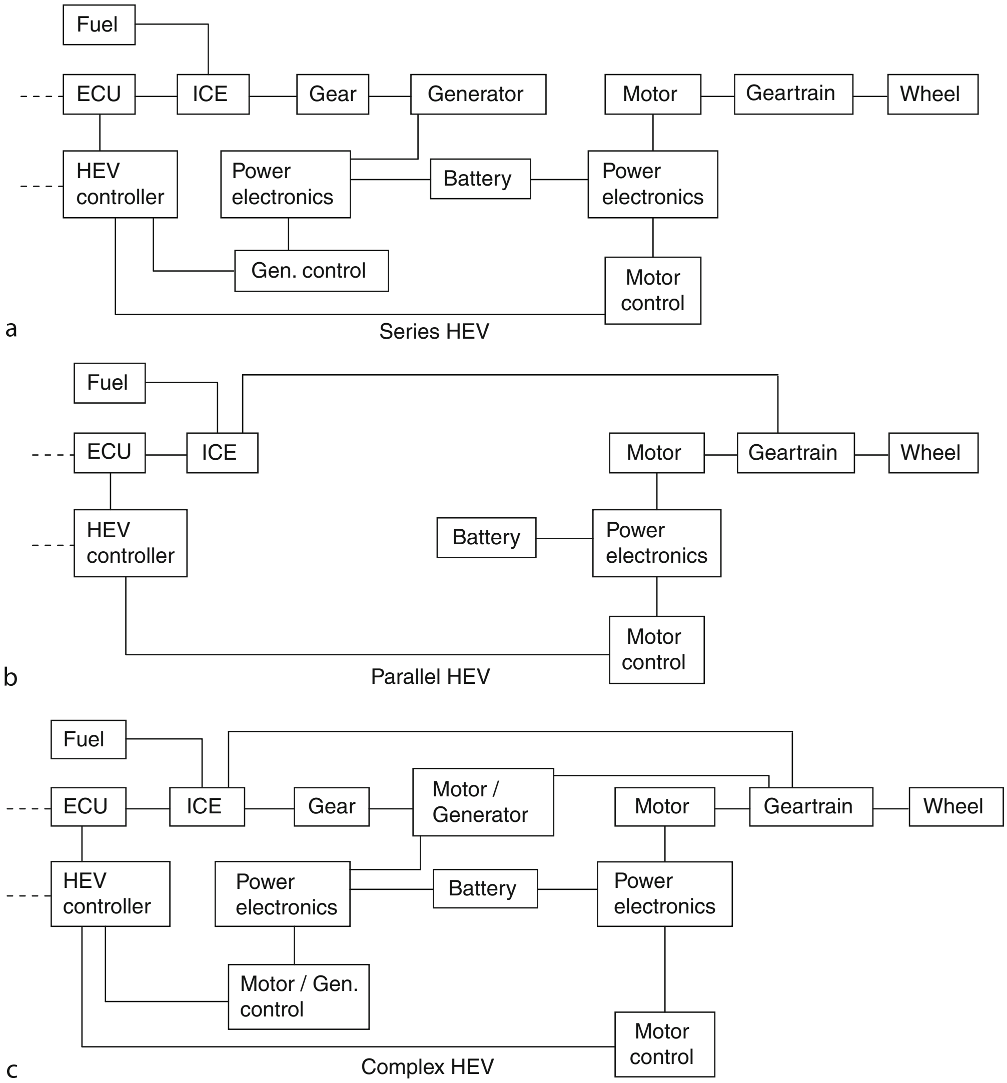
One of the most important reasons for considering HEV for vehicular applications is the cost of the fuel, reduced emissions, and acoustic noise. Several different architectures have been implemented in HEV, depending on the application. For example, a series HEV (SHEV) architecture can be implemented by using several hub motors, with one motor on each wheel, or alternatively, there could be one motor per axle for propulsion. This can take advantage of redundancy in case one of the motors fails. In this case, the vehicle can run in a gracefully degradable mode, with somewhat lesser performance and drive to safety as needed. A brief introduction of architectures used in vehicular applications is given in the following section.

Series, Parallel, and Complex Architectures for Ground Vehicle Applications

For the sake of continuity, various types of automotive HEV architectures are discussed in this section briefly. HEV architectures used in off-road vehicles are similar to automotive HEV architectures, except the engine technology. Automotive HEVs use gasoline IC engines in general, while off-road vehicles in general use diesel-based engines. (In this chapter, IC engine implies regular automotive gasoline engine unless explicitly indicated otherwise.) A very informative research paper [1] has made a comparison between the series and parallel HEV architectures (Fig. 2a, b) with a regular IC engine vehicle. From the reference cited above [1], it has been found that IC engine vehicles are slightly lighter than



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 1
A generic hybrid power system



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 2
(**a**) Series HEV configuration, (**b**) parallel HEV configuration, (**c**) complex HEV configuration (Courtesy: Pubs-Permissions@ieee.org)

parallel HEVs. However, the series hybrid vehicle was found to be relatively heavier. If a nonhybrid IC engine vehicle is considered a baseline, then one can compare the other architectures as follows. To match the performance of the baseline vehicle, obviously it is necessary to make the electric motor in a Series HEV (SHEV) to have the same size as the IC engine. The authors believe that comparison of vehicle weight and volume of an IC engine vehicle and "series HEV" should be based on drive cycle, and the average energy efficiency of the propulsion system, that is, the comparison should be done under similar circumstances in terms of its operation and type of engine used.

The peak power can be handled by the battery, which should be charged within its bounds. The authors recommend simulation and experimental studies of different drive cycles and then coming up with a worst case scenario to help decision making. The battery and other storage devices play an important role in decision making, depending upon whether they can provide the maximum power needed or not. Only if the battery is kept under completely floating condition all the time, and the power from the ICE and generator is fed to the motor in parallel with the battery, will require the generator to be sized equal to the maximum power demand. The intent of the battery or any peak power source is to address the maximum power demand in SHEV, and hence ICE and generator size in SHEV could be reduced. However, the size of ICE and the generator in SHEV will definitely depend on the average power demand by the vehicle based over the expected drive cycle in which it will operate. In a parallel HEV, the size of the electric motor could be less than half the size of SHEV motor.

The reference [1] assigns the balance of power to the ICE in the Parallel HEV (PHEV). The mechanism to split the power assignment to the motor and the ICE in the parallel HEV assumes that the original ICE in the baseline vehicle was chosen based on the maximum power needed under the worst possible drive cycle scenario. There is also no separate generator in the PHEV in the reference cited above. However, in some architectures, like Toyota Prius and some others as well, there is a motor/generator that is separate from the main propulsion motor, which can be used as generator and starter as well.

The reason for having this generator is highly dependent on the strategy for control. Under certain conditions of the battery state of charge (SOC) and power demand in the propulsion, it may be necessary to provide additional charge to the battery, through this generator.

The battery, or any other storage elements like the ultracapacitor, is an important part of the SHEV or PHEV. In specifying batteries, normally voltage and amp-hrs are used as metric. But in PHEV and SHEV applications, it is also necessary to know the current (or equivalent power) rating to satisfy the maximum current (or equivalent power) demand by the propulsion motor. The current (and power) demand from the propulsion motor can be met by the generator and the battery working together. As noted earlier, quite often the battery is used to meet some high transient (or peak) current demand from the motor. Hence, it is very important to know the battery current rating, both continuous and instantaneous, in addition to its voltage and amp-hrs. All these issues require a vital decision-making process in developing the design specifications for an HEV. The discussion above applies to both gasoline- and diesel-based HEV for better fuel economy and other benefits, such as reduced emissions.

Another question that requires a serious discussion is the choice of HEV architecture, that is, whether it should be series, parallel, or complex (Fig. 2a–c). The question can be best answered depending on what is the priority – is it fuel economy, performance, size or weight, reliability, or emissions, which are briefly answered below.

- If fuel economy is of utmost priority, then it should be a PHEV.
- On the other hand, if the performance of the vehicle is important (i.e., power output), then SHEV is a better choice. This might help reduce transmission package as well as it makes the system capable of providing high power on demand quickly. This is due to the fact that motor reacts to control the commands faster than the ICE.
- If size and lightweight are of concern due to the vehicle requirements, then PHEV will be a better choice than SHEV. In general, SHEV is to some extent heavier than PHEV, hence this fact might also affect fuel economy slightly.
- Reliability is not to be underestimated in both PHEV and SHEV. However, reliability considerations might be better addressed in general by using PHEV, because of redundancy in propulsion due to ICE and the motor. However, PHEV control system is more complex, in addition to complexity in its mechanical coupling. A thorough study of failure modes of these devices is needed before one can conclusively make a decision on reliability matters.
- Another issue accompanying reliability in vehicular systems is maintenance. It appears that maintenance of an SHEV is somewhat simpler due to several reasons. Its control and mechanical linkage, unlike PHEV, are simple. If there are hub motors in its propulsion, they can be rather quickly replaced in case of failure.
- Off-road and military applications of these vehicles require device ruggedness. This implies that both mechanical and electrical as well as other design

aspects like high temperature extremes (both hot and cold) must be considered. In general, high temperature and mechanical vibrations can negatively impact items like power electronics and batteries. This means the devices can get physically larger and bulky to provide for appropriate cooling needs, if performance is to be achieved at these extreme conditions.

- The above discussion indicates that there are pros and cons of PHEV and SHEV. Having said that, it should be noted that the decision regarding the choice between PHEV and SHEV could be very much application and drive cycle dependent. Final decision on the above requirements is quite complex, and systematic trade-off studies are required for all requirements like performance, fuel economy, reliability, etc., to reach most optimum decision.

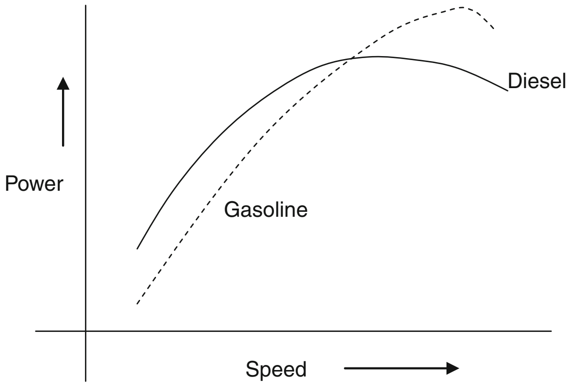
Off-Road Vehicles

Success in hybrid electric vehicles in automotive applications has led certain other vehicular areas to consider this technology as well. These include ships and aircrafts, which are nonground vehicles. Diesel electric locomotives have already been using the technology in a slightly different form and these will be discussed as well. Off-road vehicles are ground vehicles for heavy-duty applications. These vehicles, in general, do not use normal roads for their operation. This entry specifically focuses on heavy-duty applications of off-road vehicles. Examples of such vehicles are construction vehicles and equipments (e.g., mining vehicles/trucks, refuse trucks, agricultural vehicles like tractors, etc.) and some military vehicles. Off-road vehicles, in general, operate at low speeds compared to regular passenger vehicles and use diesel engines, instead of gasoline IC engines. Use of diesel engines has certain advantages compared to gasoline IC engines as discussed in the following.

- The compression ratio of diesel engine can be much higher than the gasoline, for example, it can be around 15–20% compared to 9–12% for the gasoline engine. The higher compression ratio necessitates the engine to be heavier, which helps in construction to sustain the cylinder pressure.
- At low speeds diesel engines are more efficient than gasoline IC engines and are better in terms of fuel economy due to the absence of throttle vales, and avoid parasitic losses. This leads to smaller-sized (HP ratings) engines, specifically at lower speeds.
- Other benefits of diesel engines include lower greenhouse gas emissions such as CO_2.
- Low speed characteristics of diesel engines are better compared to gasoline engine, in terms of starting torque generation.
- Diesel engine has less components due to the absence of ignition system, and hence leads to higher overall system reliability.

A comparison of gasoline and diesel engine speed–power characteristics is shown in Fig. 3. It is apparent that at lower speed range diesel engine can provide higher power compared to a gasoline engine of comparable size. This makes it suitable for high-power low-speed type of operation, which is more consistent with off-road vehicles.

The drive cycles and speed-torque demands of these off-road vehicles can be quite different as compared to regular automobiles. Off-road mining/trucks and construction vehicles operate in rough terrain and at low speeds or under stationary conditions. Therefore, high power and torque are required due to the nature of these vehicle operations. In order to supply high torque and power requirements from a diesel engine, it is necessary to have a large gear box and transmission system, etc. In addition, the overall efficiency of diesel engine is still low (of the order of 40%, but still may be



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 3
Comparison of gasoline and diesel engine curves

higher than gasoline engine) over the complete drive cycles (or torque speed demand of the load).

To alleviate these issues, hydraulic or electric hybrid propulsion systems are being developed and have been used lately. Hydraulic systems operate by continuously running the engine, which is a low-torque high-speed system relative to load. The engine is used to drive a hydraulic pump which can drive a hydraulic motor. Through the use of an incompressible fluid one can achieve very high pressure, which can be applied at the load drive end. However, the hydraulic systems, in general, can cause problems. Examples of such problems are complex plumbing system with pipes, which, under rough terrain conditions, can cause plumbing system failures. This may be due to fluid leaks, leading to degraded performance and eventually hydraulic system failure. Other disadvantage of hydraulic systems is the parasitic losses.

Based on the above discussion of diesel engines and their use in off-road vehicles, it can be concluded that the merits of diesel engine combined with electric propulsion (diesel HEV) can lead to most fuel-efficient vehicles, subject to the specific application needs. Hence, for the same performance, and particularly during acceleration, it is beneficial to use diesel hybrid, within the same package size. This is especially beneficial in a parallel configuration, when both the engine and the electric propulsion are required to play an important role to achieve optimum performance in terms of power and acceleration. The issue of performance is inherently better addressed in series hybrids, since propulsion is done only by electric motor.

In HEV, energy storage and electric propulsion motor systems offer very high efficiency of the order of 75% and above. In present HEV, engine is used to charge battery and the propulsion torque/power can be shared between the electric propulsion and the engine. Hence, lately various heavy off-road vehicle manufacturers of mining and construction equipments are paying serious attention to transform their system to electric hybrid type of propulsion. These systems offer clean environmental advantages in addition to the fuel economy. It must be noted that hybrid electric off-road vehicles (HEV) are different than the regular automotive HEV, and do not use normal roads for their operation. Mining vehicles, agricultural vehicles like tractors, and some military vehicles also fall in this category. Off-road vehicles will be discussed further in the following section, which have quite different drive cycles and speed-torque demands, as opposed to a regular automobile.

Off-Road HEV

In the preceding sections off-road heavy vehicular applications like mining and construction were mentioned. These vehicles require large transmission system when equipped with only a diesel engine drive, which has relatively low overall efficiency. These issues can be circumvented through HEV system having overall efficiencies of the order of 75%. In an off-road HEV a diesel engine is used to drive a generator or alternator, which in general will generate a variable speed (hence variable frequency) variable voltage (amplitude) system. This variable frequency and variable voltage is then converted to a DC voltage of constant value using an AC to DC converter/rectifier. The constant DC voltage is then converted to a three-phase AC system (inverter) to drive a three-phase AC motor. The inverter can be electronically controlled to deliver desired three-phase voltage (amplitude, frequency, and phase) to drive an electric motor to handle the vehicle load. It should be noted that in the system just described, one can avoid the use of a battery. Although battery is a vital storage element in a regular HEV, in very large off-road vehicles, the battery system can be avoided to reduce cost and package size. In addition, in certain types of vehicles sufficient energy is not available for storage during the vehicle operation. Such a system is very similar to a diesel electric locomotive system.

Various architectures of the off-road and other construction vehicle systems are possible. Two possible systems are shown below. In this architecture the IC diesel engine, in general, is operated at its optimal speed to achieve maximum efficiency. IC engine drives an alternator/generator, which generates a variable three-phase AC voltage (amplitude and frequency). This is rectified and a DC is produced. The DC power is then fed to the inverter to produce desired voltage (amplitude and frequency) to drive a propulsion motor. In general, for very heavy mining applications, induction motors are found to be more robust and suitable in rough terrain/environment. The motor is then used to drive the vehicle wheels. Generally, large off-road vehicles have two axles and two motors (one on each axle) for propulsion.

In the architecture shown in Fig. 4, the use of battery is avoided. In principle, a battery could be used for energy storage in vehicles, but the size of the battery in that case would be large. In some applications an ultracapacitor bank can also be used. The mechanical energy, during slowing down of the vehicle or some of its movable parts, can be stored in the ultracapacitor as electrical energy. The ultracaps normally have high specific power and low specific energy, that is, it can provide very high burst of power input to a device for a short duration. During the slow down, the controller commands the ultracapacitor, inverter, and the alternator/rectifier to coordinate properly, so that energy flows through the motor and inverter (with the motor operating as a generator) into the ultracapacitor. The stored energy can be utilized to accelerate the vehicle, if so needed at a later time, to save energy and thus leading to better system efficiency and fuel economy. An architecture for such a system noted above is shown in Fig. 5. In this architecture to drive the motor, the DC power could be provided by the ultracapacitor or the IC engine/rectifier. The system will have the necessary switches to select either the ultracapacitor or the alternator/rectifier or both.

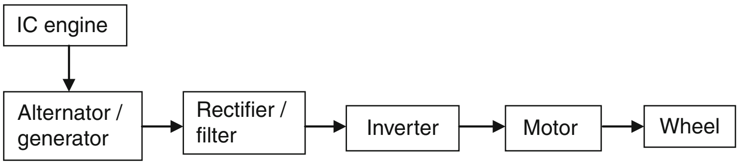
It will be interesting to visit some typical sizes involved in such off-road vehicles discussed above. For example [2–4]:

- A Caterpillar engine model Cat 3524B EUI has gross power 2,648 kW (3,550 HP), machine gross operating weight 623,690 kg (1,375,000 lb).
- A particular Komatsu has gross horsepower 2,611 kW (3,500 HP), and weight of 505,755 kg (1,115,000 lb).
- A particular Liebherr has gross horsepower @ 1,800 rpm of 722 kW (3,650 HP), 592 t (British)/ 652.5 t (US).

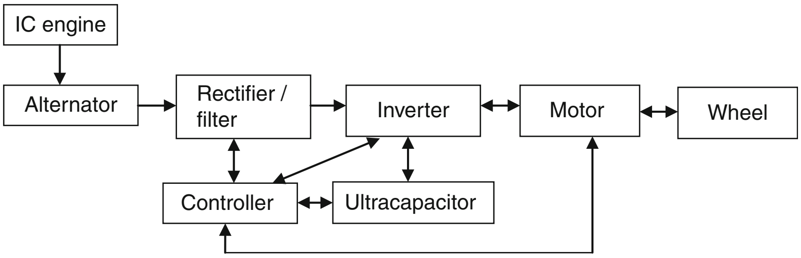
The sizes of the tires of these vehicles are in general of the order of 4 m, as is evident from the specifications. This indicates the huge size of these vehicles, some pictures of which are given in Fig. 6.

All the above vehicles, that is, Caterpillar, Komatsu, and Liebherr, are transitioning to hybrid (HEV) power train. Appendix 1 provides additional specifications from the Komatsu information sheet [3] for vehicle shown in Fig. 6b.

The traction motors in both Komatsu and Liebherr are induction motors. The alternators are normally brushless. The gear ratio, reduction from motor to wheel, is of the order of 32–37. The gross engine horsepower of the vehicles is about 2.6–2.7 MW. The inverter



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 4
System-level architecture of a diesel electric HEV



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 5
System-level architecture of a diesel electric HEV with ultracapacitor

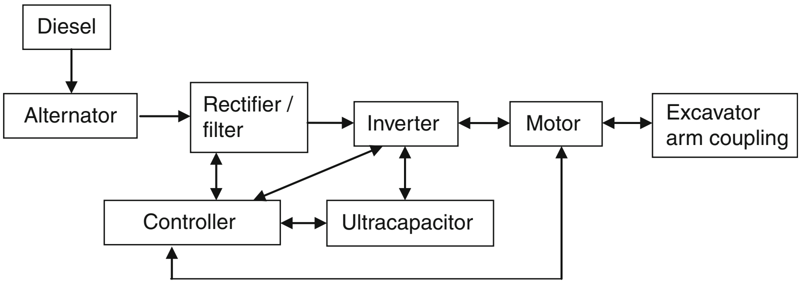
Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 6
Typical mining vehicles: (**a**) Caterpillar, (**b**) Komatsu, (**c**) Liebherr (Courtesy: Caterpillar http://catsays.blogspot.com/2005/01/caterpillar-797b-mining-truck.html, http://www.cat.com/contact. Komatsu http://www.komatsuamerica.com/default.aspx?p=equipment&f1=view&prdt_id=947, http://www.komatsuamerica.com/contact. Wikimedia http://commons.wikimedia.org/wiki/File:Liebherr_T282.jpg [2–4])

used to drive the motor generally uses IGBT for switches. Assuming alternator and inverter efficiencies of the order of 0.95 each, the inverter power output (or motor power input) is about 2.35–2.44 MW. So it is reasonable to assume that each induction motor size has to be of the order of 1.2–1.25 MW (assuming two motors for the system, with one at each axle). The vehicle at maximum speed amounts to a wheel revolution of about 84 rpm (assuming a tire diameter of about 4 m). With a gear ratio of about 32.62 the motor speed should be about 2,700 rpm. Nominally, this leads to about 3,000 rpm, which, at 50 Hz, leads to a four-pole induction motor. Four-pole induction motor is normally a common standard for such applications.

Another application of HEV in current off-road vehicles involves construction vehicles such as excavators and scrapers (also known as "track"-type tractor). It is expected that an array of other off-road vehicles may transition to HEV technology, once the customers start accepting the advantages offered by this technology.

Presently in excavators the usage of HEV concept is used in its arm movement, and is not used in vehicle propulsion. The excavator arms need brief burst of high power followed by low power return and then slow down its speed to stop the arm. Such an application is very well catered by using a diesel engine driving a generator to produce electricity to drive the excavator arm motor, which is an electric drive. The possible architecture is shown in Fig. 7.

This architecture is basically the same as the HEV architecture shown previously for the mining truck



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 7
System-level architecture of an excavator with hybrid electric technology and using ultracapacitor

(or hauler) propulsion. Only difference in this case is that the propulsion size is much smaller. It has been claimed that the Komatsu hybrid excavator leads to an additional fuel economy of order of 25% and that some users have achieved fuel economy enhancement as high as 41% for specific applications. A picture of the Komatsu PC 200–8 hybrid excavator is shown in Fig. 8 [5].

Presently, excavators are considered to be good candidates for the installation of energy storage devices such as battery or ultracapacitors. This is perhaps due to size of the machines which require smaller and lower cost storage devices as compared to mining trucks. Some manufacturers prefer to use battery instead of ultracapacitor for the purpose of storage. An example is the New Holland's excavator, which uses a 36 HP diesel engine, 20 kW generator, and a 288 V lithium battery. The choice of storage is highly dependent on exact usage and duty cycle. If it is just a high-power excavator, then ultracapacitor seems to be a suitable candidate. However, if in addition to excavator application, there are other auxiliary equipments, which call for energy that may not be immediately available from an ultracapacitor, a battery may be used. The duty cycle in terms of torque requirement, duration of the torque, auxiliary equipments, cost, size, and similar factors all together leads to the decision process in terms of what will be best, that is, ultracapacitor or battery.

Track-Type Tractor Scraper/Bulldozer Another application of off-road construction and agricultural equipment is a bulldozer or scraper. This was first introduced by Caterpillar in late 2009 [11] and should be available in 2010. The structure of the vehicle, the D7E model is shown in Figs. 9 and 10 [11, 12].

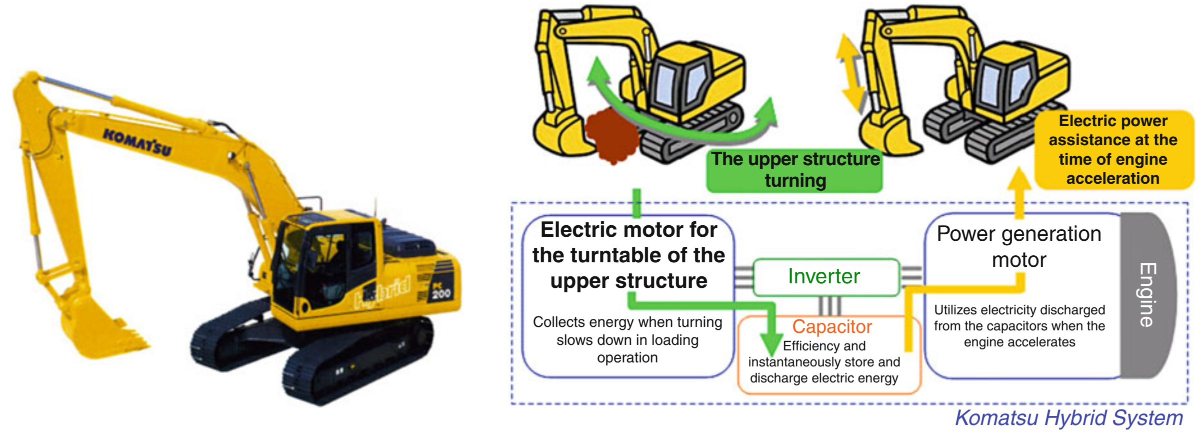
The propulsion is achieved by a diesel electric system with a net power of 175 kW. Here a diesel engine drives a generator. The AC generator and propulsion module voltage is 480 V. The electrical power from the generator is processed by an inverter, which drives a motor. D7E has three-phase induction motors for its drive [13]. It has been indicated [11] that the cost of this model is 20% more than its previous D6 and D8 models. This 20%, which is about $100,000, is claimed to be recoverable in 2 years due to fuel savings of about 10–30%. Performance of this is also claimed to be better than its previous versions. While gas mileage is of importance, in this type of machines the efficiency is measured through how much material can be moved per gallon of fuel, because these machines are not predominantly vehicles in the sense of traveling distances. In terms of materials moved, it is apparently 10–30% more efficient than its nonhybrid equivalent. This machine does not have battery for energy storage, and the power from the generator is directly processed by the inverter and transferred to the motor. The machine can apparently save about 6 gal of fuel over an 8-h period, which can be significant if it is used continuously throughout the year.

Off-Road Military Vehicles

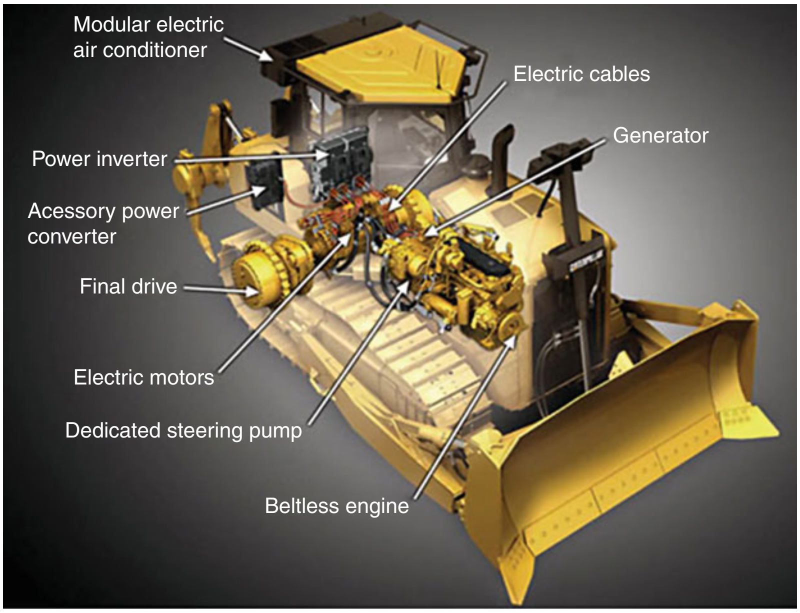
In addition to commercial off-road vehicles, military application also includes similar vehicles. These vehicles, in principle, have similar requirements.

Several pictures of off-road military vehicles are shown in Fig. 11 as illustration and to help readers get visual perspectives of these machines [14–20].

Figures above show vehicles which are regular propulsion type, and could be enhanced by HEV technology.


KOMATSU
Hybrid
PC 200
The upper structure turning
Electric power assistance at the time of engine acceleration
Electric motor for the turntable of the upper structure
Collects energy when turning slows down in loading operation
Inverter
Capacitor
Efficiency and instantaneously store and discharge electric energy
Power generation motor
Utilizes electricity discharged from the capacitors when the engine accelerates
Engine
Komatsu Hybrid System

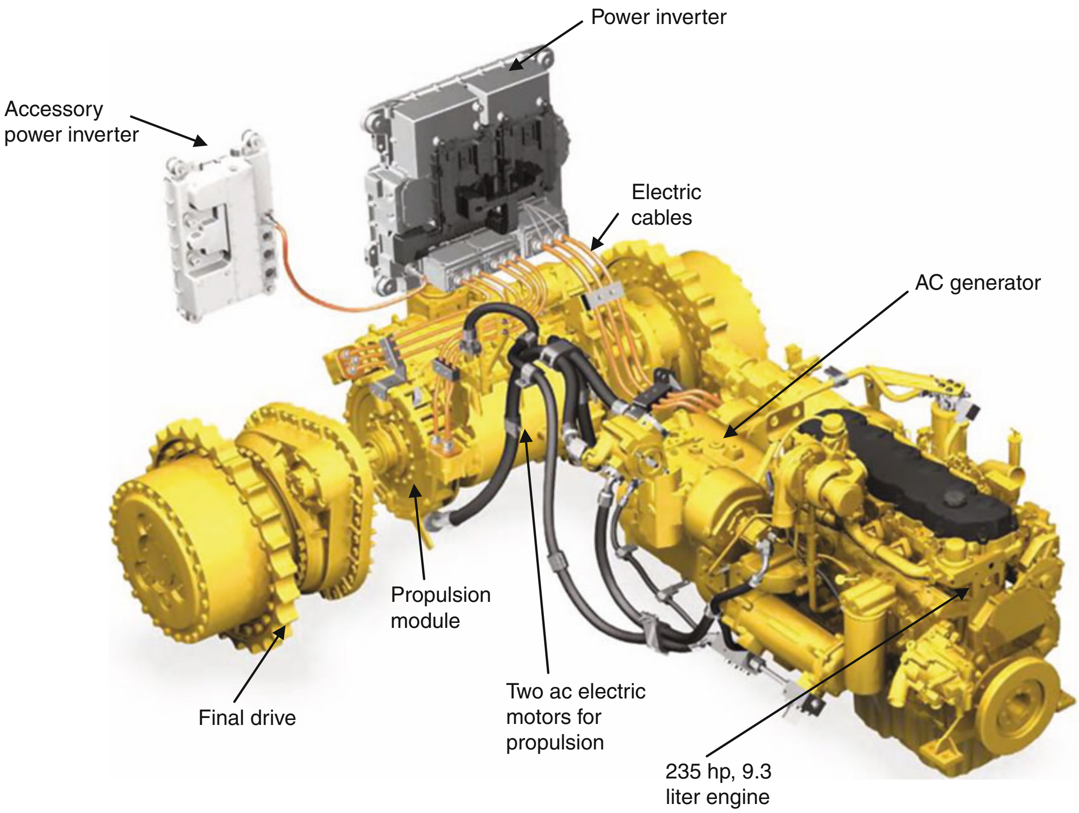
Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 8
Komatsu excavator system (Courtesy: Komatsu news item http://www.komatsu.com/CompanyInfo/press/2008051315113604588.html [5])



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 9
Caterpillar D7E propulsion system (Courtesy: Caterpillar Information Brochure Web article, http://gas2.org/2009/09/18/caterpillar-builds-worlds-first-hybrid-bulldozer/, Web article, http://www.cleanmpg.com/forums/showthread.php?t=33855, http://www.cat.com/contact [11, 12])

From the various types of vehicles shown in the pictures above, it must be noted that these vehicles have different operating requirements in terms of their drive cycles and power when compared to regular automotive applications.

Some vehicles run long distances, and some have extreme power demand and performance needs, such as survivability and fuel economy. Introduction of hybrid propulsion can lead to fuel economy, which has direct and indirect benefits. Direct benefits come



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 10
Details of the diesel electric power train organization in Caterpillar D7E (Courtesy: Caterpillar Information Brochure Web article, http://gas2.org/2009/09/18/caterpillar-builds-worlds-first-hybrid-bulldozer/, Web article, http://www.cleanmpg.com/forums/showthread.php?t=33855, http://www.cat.com/contact [11, 12])

in the form of fuel economy, which leads to cost savings. However, for military applications, carrying of fuel to the field involves additional cost and need for logistic vehicles; hence the impact of fuel economy is even higher, compared to commercial vehicles. Hybridization can also lead to redundancy (particularly for parallel architecture with both IC engine and electric propulsion), which can have positive impact on vehicle survivability in case of partial failure. It is also important to note that military environment can involve a wide temperature range and high NVH (noise, vibration, harshness). Power electronics and electric motors have to withstand such situations. These are important enabling technologies which have to undergo further transformation to make HEV technology to gain wide acceptance in military applications.

Summary of Off-Road HEV

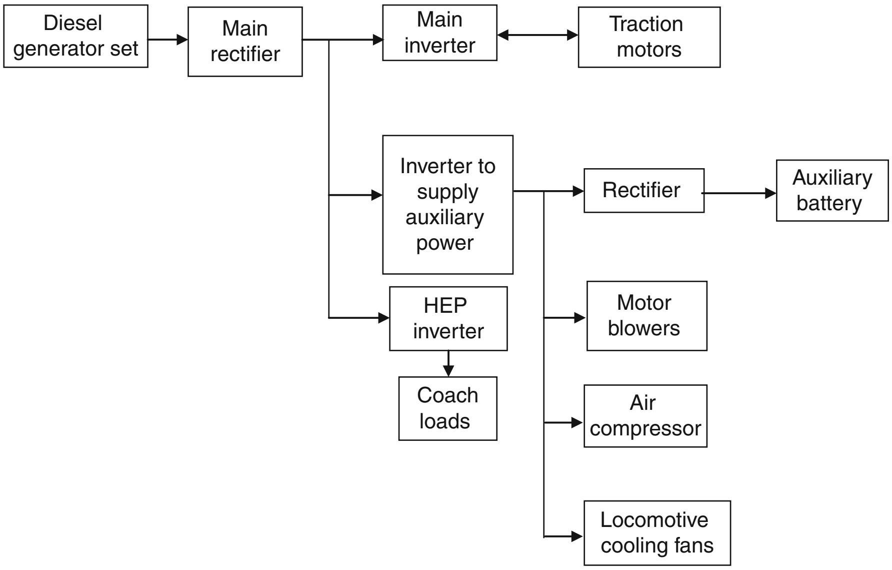
Regarding the rating of the motor for the mining truck – nominal DC link rating of these motors can be around 2,600 V, leading to about 1,600 V line-to-line rms input to the three-phase motor. A 1.2 MW motor corresponds to about 541 A AC rms, assuming a power factor of about 0.8. In this case the DC link current for just one motor will be 1.2 MW/2,600 = 461 A DC per motor. All these are assuming no additional loss in the system and the numbers are merely giving an idea of the voltage and current values. Some systems [6] use 3.3 kV IGBT technology to drive these motors, which is consistent with the above ratings indicated for the motor. For the excavator example above the numbers are much lower. As per preceeding paragraph, at 288 V DC, the DC link current will be about 70 A,

Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 11
(**a**) Stryker (http://commons.wikimedia.org/wiki/File:14_stryker.jpg) (**b**) HEMTT (http://commons.wikimedia.org/wiki/File:HEMTT.jpg) (**c**) Bradley (http://commons.wikimedia.org/wiki/File:Two_M-3_Bradleys.jpg) (**d**) Abrams (http://commons.wikimedia.org/wiki/File:M1A1_Abrams_Tank_in_Camp_Fallujah.JPEG) (**e**) Fennek (European) (http://commons.wikimedia.org/wiki/File:BundeswehrFennek.jpg) (**f**) MRAP (http://commons.wikimedia.org/wiki/File:FBI_Mine_Resistant_Ambush_vehicle.jpg) (**g**) Gladiator (http://commons.wikimedia.org/wiki/File:Gladiator_RSTA.jpg)
Courtesy of the websites indicated in the above references [14–20]

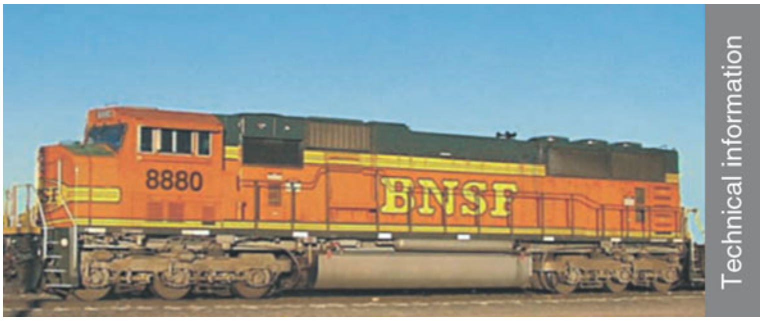
the three-phase line-to-line rms for the motor will be about 0.612 × 288 = 176 V, motor current will be about 82 A.

Diesel Electric Locomotives

Diesel engine–based locomotives are commonly used nowadays. In diesel electric locomotives the propulsion system is electric as explained through the block diagram in Fig. 12. There are advantages why diesel electric system is used, as explained in sections "Introduction" and "Series, Parallel, and Complex Architectures for Ground Vehicle Applications" (i.e., if the propulsion was purely mechanical, then a big-sized transmission system with gearbox etc. will be needed to create the needed torque). A second advantage in using electric propulsion system in locomotives is that the electric propulsion system provides continuously variable speed system. This is particularly advantageous in improving the transmission efficiency of short haul trains where speed fluctuations are encountered frequently. Otherwise, a diesel engine with a finite number of gears is required to be operated at different speeds other than the most optimum speed at which the efficiency is highest. In this system (HEV), the diesel engine in general runs



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 12
System-level architecture of a diesel electric locomotive



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 13
A large diesel electric locomotive (Courtesy: Siemens Technical Information Brochure http://www.siemens.cz/extra/msv/cd/TS/Vozidla/Diesel-Electric_Loco_SD70MAC_EN.pdf, contact.mobility@siemens.com [7])

at the most optimal speed and drives the generator to produce electricity, which is used to run a traction motor mechanically connected to run the wheels as shown in Fig. 9. With the advent of power electronics, this system is rather easy to realize, using reliable and efficient traction motors, which can be induction or synchronous motor – either permanent magnet or field-excited types. With this background, let us look at the architecture of the diesel electric system shown in Fig. 12.

In general, the traction motors are placed on each axle of the locomotive to drive the wheel pairs. These traction motors used in heavy locomotives are AC induction motors. The squirrel cage induction motors are a very reliable device for such applications, due to their robust designs and structures and resilient to temperature conditions, unlike permanent motors.

A picture of a diesel electric locomotive is shown in Fig. 13 [7].

Its diesel engine is rated at 4,000 HP and power at wheels is 3,350 HP. Six motors are used in this engine, with one per axle. The motors are four-pole squirrel cage three-phase induction motors, with a maximum voltage of 2,030 V and 433 kW continuous power rating. The net motor power rating is 2,598 kW, or about 3,500 HP. The maximum speed of the motor is 3,220 rpm, and there is a gear ratio of 85:16 between the motor and the wheel. Although specific power density of induction motors can be somewhat lower than permanent magnet motors, in locomotive applications, where the system package size may not be an important consideration (as in small passenger cars), this may be an ideal choice. As shown in block diagrams, in general, no traction system batteries are presently used in locomotive applications. To start the diesel engine, either electric, hydraulic, or pneumatic starting mechanism can be used. In case of electrical start, a small starter battery is needed.

Recent advances in technology, for capturing regenerative energy, involve some storage battery, ultracapacitor, or a combination of both. Even a flywheel storage unit, can be considered for regeneration. Regeneration can help improve fuel economy and is of more importance for short haul trains and may not be as important for a long haul train.

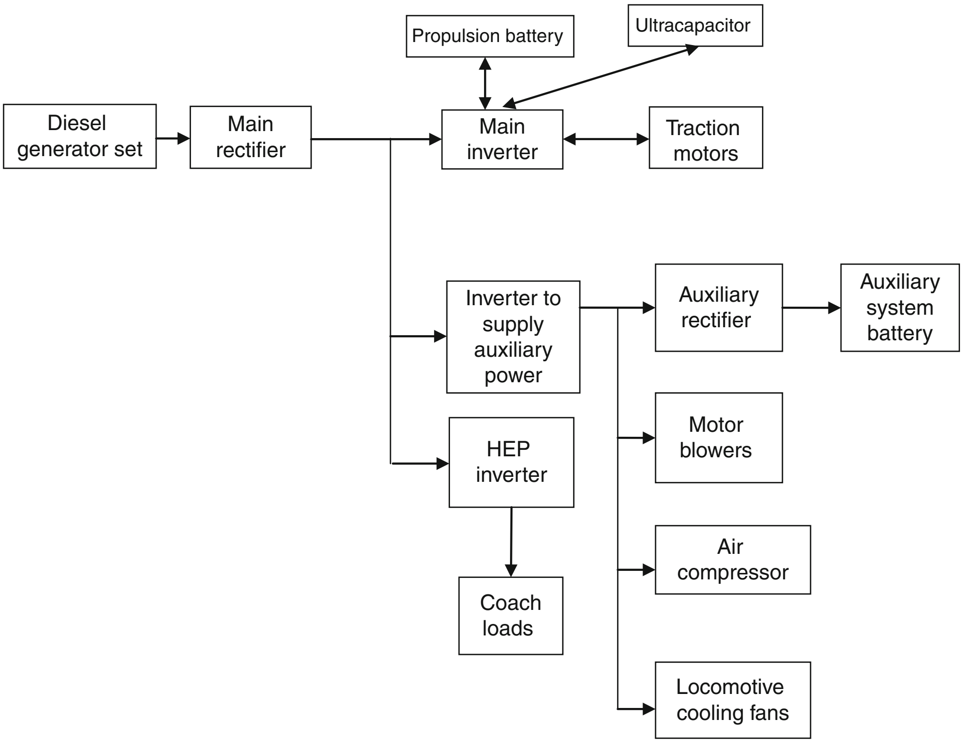
The architecture of a locomotive system capable of regeneration is shown in Fig. 14. The system is essentially the same in principle as the one in Fig. 12, except that it has a propulsion battery and/or ultracapacitor, which can provide energy to the propulsion motor.

Hybrid Hydraulic Vehicles

Nonelectric hybrid vehicles, specifically *hybrid hydraulic vehicles* (HHV), are being considered by the technical community as viable and can sometimes be more beneficial than an electric hybrid vehicle.

Energy Storage

As compared to electric energy storage system, for example, battery in HEV, in a hydraulic system the



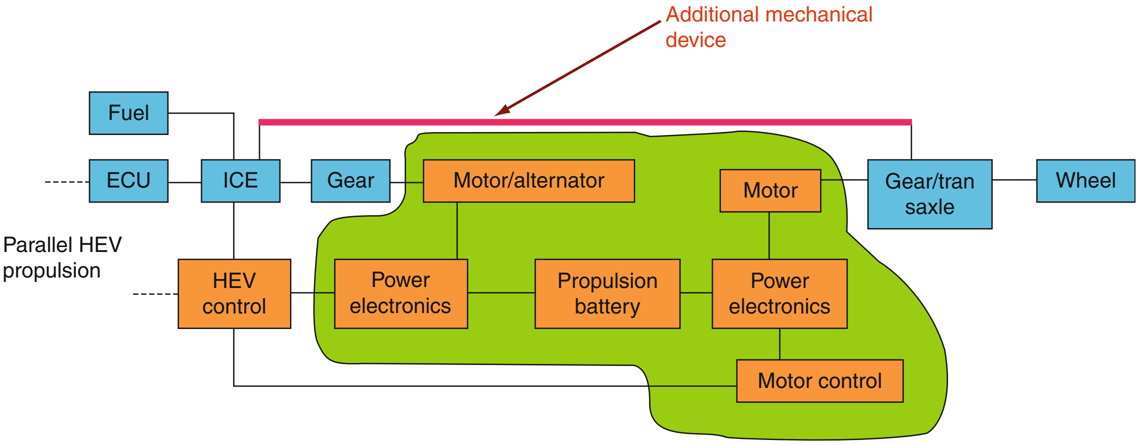
Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 14
System-level architecture of a diesel electric locomotive including the usage of ultracapacitor

energy is stored in the form of compressed gas in a cylinder. To pressurize a gas, one needs power and energy, which comes from the IC engine to activate the hydraulic pump. While extracting the energy one can use a hydraulic motor. In other words, the hydraulic pump is analogous to an electrical generator, since the latter converts mechanical power into electrical, whereas hydraulic pump converts mechanical power into fluid power. Similarly, hydraulic motor is analogous to an electric motor, and the pressurized gas in the cylinder is analogous to a battery. The efficiency of the hydraulic pump, motor, and storage is around 90%.

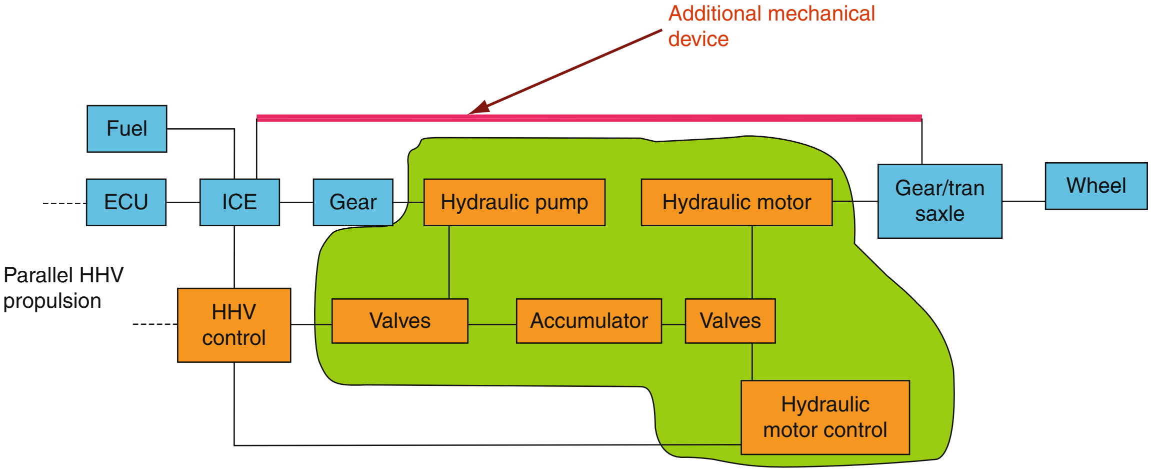
To illustrate the hydraulic hybrid system, one can compare Figs. 15 and 16.

For the hybrid hydraulic vehicle architecture shown above, the correspondence with its electrical system counterpart is shown in Table 1.

Figure 17 shows a full series hybrid hydraulic power train configuration [8], which corresponds with the system-level diagram shown in Fig. 16. This includes a high-pressure (HP) accumulator (Appendix 2) containing benign gas such as nitrogen, which can be compressed or expanded by means of incompressible hydraulic fluid acting through some piston type of



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 15
System-level diagram of HEV



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 16
System-level diagram of HHV

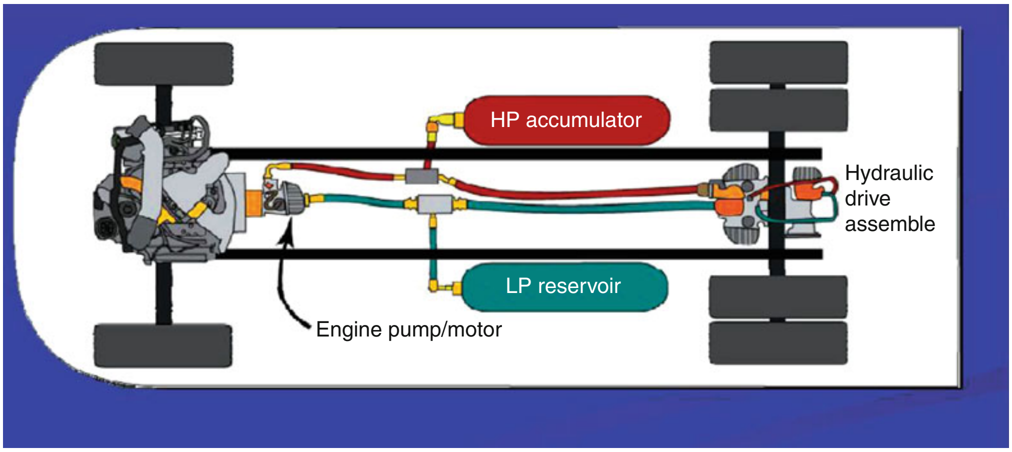
mechanism. The pressure in HP cylinder and the pipeline can be as high as 3,000–5,000 psi, whereas the low-pressure cylinder and the corresponding pipeline containing incompressible fluid can be of the order of a few hundred psi. The IC engine drives a pump, which converts the pressure from low to high as the fluid is moved from LP to HP cylinder. To drive the vehicle wheels, the high-pressure gas from the HP cylinder may be allowed to expand and the hydraulic fluid is moved through the motor drive assembly. The hydraulic motor takes the high-pressure fluid in its input port, and converts it to mechanical power at the wheels. When the fluid has passed through the hydraulic motor, its pressure drops and the low-pressure fluid is transferred to the LP cylinder.

The amount of energy stored in the hydraulic accumulator system is quite low in terms of kW-h/kg. As a comparison, the specific energy density in hydraulic accumulator can be of the order of 1.9 w-h/kg [Ref EPA], whereas a battery can have specific energy density of 30 w-h/kg. However, the specific power density of hydraulic system can be of the order of 2,500 W/kg, whereas the electrical system–specific power density (e.g., a battery) can be about 650 W/kg. Therefore, the hydraulic hybrid system can be more suitable for high-power and relatively low-energy applications, particularly where short bursts of high-power acceleration and deceleration are involved.

It is evident from the earlier discussions that an IC engine can run at optimum efficiency to drive the hybrid hydraulic system to deliver the energy to the wheels. A few scenarios can take place here. If the energy needed by the vehicle propulsion matches the energy from the IC engine, the power generated will be used effectively in propulsion. However, if the energy coming from the IC engine is greater than the propulsion load, the excess energy will go to the accumulator for storage. The energy stored in the hydraulic accumulator is rather low, hence this option cannot continue too long, since excessive gas pressure in the accumulator will then be needed, which can be difficult

Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Table 1 Equivalence of major components between hybrid electric and hybrid hydraulic vehicular systems

Electrical system component	Hydraulic system component
Alternator	Hydraulic pump
Electric motor	Hydraulic motor
Battery	Hydraulic accumulator
HEV controller	Hybrid hydraulic vehicle (HHV) controller
Power electronic system	Hydraulic valve system

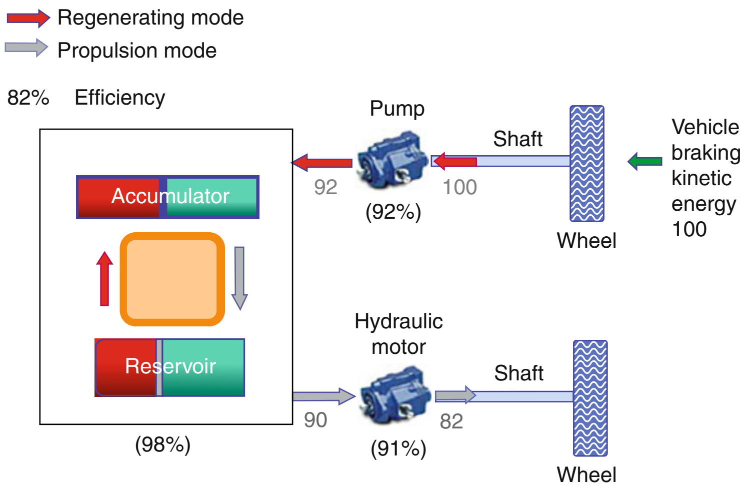


Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 17 Series HHV system architecture (Courtesy: EPA Kargul, J.J. “Hydraulic hybrids – demonstration for port yard hostlers,” in EPA presentation, www.epa.gov/otaq/technology, July 11, 2007 [8]) to sustain. If the HP accumulator cannot store any additional energy, then it will be necessary to change the engine operating point, so as to match the propulsion needs. Alternatively, it may be possible to stop the engine for a short while and use the energy from the hydraulic accumulator, and then start the engine again. As long as the average energy consumption by the motor is same as the average energy delivered by the engine, the process is sustainable, with transients absorbed by the accumulator. If there is no energy storage accumulator in the system, then in principle it is possible to keep the engine running at an optimal speed-torque point. In that case it will be necessary to switch the hydraulic valves on and off in a way such that the average fluid flow corresponds to the average power and energy used by the hydraulic motor. By following this process, it may not be necessary to perform any stop-and-go operation by the engine. But this will lead to a pulsating torque at the motor shaft output, whose average value will correspond to the average load demand.

One of the major advantages of using hydraulic system for propulsion is the fact that it can eliminate the need for a massive conventional transmission system. Also, hydraulic fluid can be more easily moved from one point to another without the need for an elaborate transmission gearbox etc. However, this concept of transmission can be used in a regular nonhybrid vehicle as well, where only the transmission system could be hydraulic. But as described in the previous paragraph, to be able to drive an IC engine at its most optimum point and at the same time eliminate torque-speed pulsations at the load will need some sort of storage, which can be an accumulator. Without a storage, it is, in principle, possible to realize the optimum operating point of the engine by switching the valves on and off to match the average load demand, but the output torque-speed at the load will then be pulsating, which can be smoothed by using a flywheel etc.

Regenerative Braking in Hybrid Hydraulic Vehicles

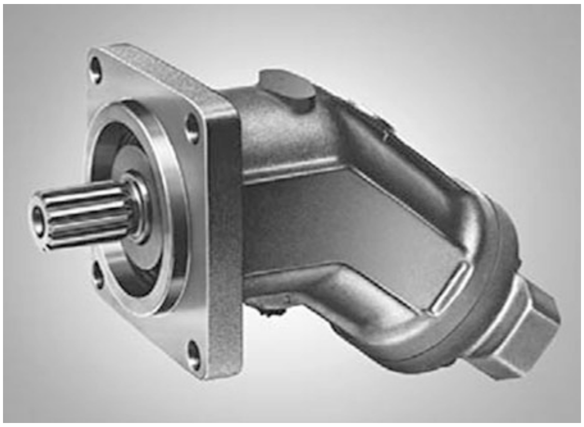
Hybrid hydraulic vehicles offer the benefit of regeneration and the ability to capture the regenerative energy (which would have been otherwise wasted) during slowing down due to braking and later use it to accelerate again. The hydraulic system–specific energy density or W-h/kg is relatively low compared to a battery. However, hydraulic system can be quite adequate for storing the regenerated energy during braking, since power is high but energy involved is typically low in this situation. Figure 18 shows the regeneration process [8].



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 18
Regenerative braking efficiency in HHV (Courtesy: EPA US EPA, “Clean Automotive Technology. . . Innovation that Works,” in www.epa.gov/otaq/technology World’s First Full Hydraulic Hybrid SUV, Presented at 2004 SAE World Congress [8])

In Fig. 18, the given efficiency of both the hydraulic pump and the motor are little over 90%. The efficiency of the accumulator (HP and LP together) is about 98%. Hence, the efficiency of the whole regeneration process is 82%.

One advantage of the hybrid hydraulic system is the maturity of the hydraulic system technology. Components used in a hydraulic system, shown in the figure above, have very high efficiency. Figure 19 shows a typical hydraulic motor/pumps (hydraulic pump and hydraulic motor are similar in appearance and their speed-torque characteristics) used in hybrid hydraulic vehicles. This figure shows a bent-axis hydraulic motor, which is from a Bosch-Rexroth specification sheet [9]. For this motor the nominal gas pressure at the entrance can be over 5,000 psi, maximum speed can be 4,500 rpm, and torque @ 5,100 psi can be 371 lb-ft. The length of the unit excluding the geared shaft is about 5.15 in.

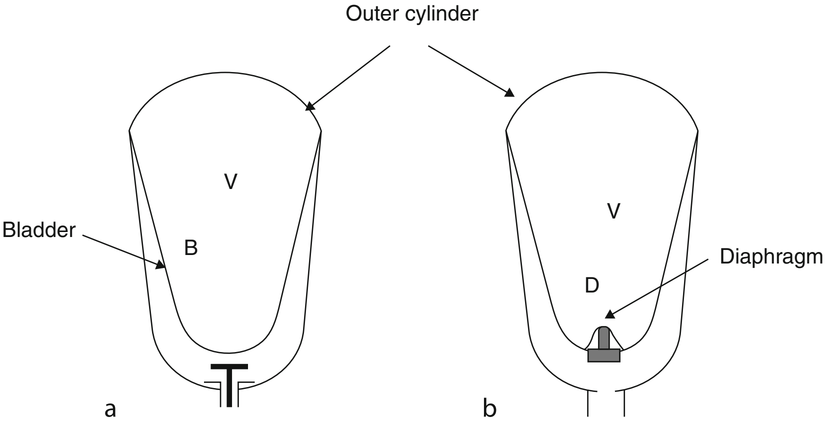


Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 19
Bent-axis hydraulic pump/motor (Bosch-Rexroth) (Courtesy: Bosch-Rexroth Rexroth – Bosch Group, "Axial Piston Fixed Displacment Motor AA2FM (A2FM)," in Product brochure from Bosch. https://www.boschrexroth.com/various/utilities/contact/index.jsp?cont_country=5&oid=3901 [9])

Future Directions

All people (technical, management, and nontechnical alike) involved with Advanced Heavy Hybrid Propulsion Systems need to work on developing and validating cost-effective, advanced next-generation heavy hybrid components and systems, which at the vehicle level could enable a power train efficiency increase as high as 100% relative to today's conventional power train technology. It is important to focus on the development and improvement of various interdisciplinary technologies, systems, and tools, such as:

- *Improved diesel engine concepts:* Advanced and novel concepts in the areas of hybrid diesel engines, fuel-injection systems, combustion- and emission-



Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles. Figure 20
(**a**) Bladder accumulator (**b**) Diaphragm accumulator

control technologies, lightweight materials, particularly for diesel engine structure, to withstand high pressure, and other novel hybrid diesel engine concepts with the potential of increased fuel efficiency and reduced emissions.

- *Low-cost, lightweight, high temperature withstandable power electronics:* Heavy hybrid vehicles will require low-cost, lightweight, simpler power electronic systems that can be easily integrated into off-road vehicular system architectures. It will also call for power electronics that can withstand very high temperature environment.
- *Energy conversion for auxiliary power and energy storage technologies and systems*: Although auxiliary systems are not used for propulsion, newer vehicles have significant amount of nonpropulsion power demand. Hence, new technologies and concepts in areas like integrated thermoelectric systems, photovoltaic, and even wind energy type of technologies are important.
- *Energy storage technology*: Battery, ultracapacitor, and flywheels need significant development to support heavy vehicle hybridization.
- *Advanced transmission and power train components and systems*: Better continuously variable transmissions (CVTs) technology and their integration into heavy hybrid applications are of importance, which can help the hybridization process.
- *Advanced materials for hybrid power trains and vehicle systems*: High-temperature materials, composites, and various polymer materials that lead to enhanced energy use and management, and reduced component and system weight and volume.
- *Rare-earth free materials*: One important area that needs to be attended to is the development of electric motor and machines technology, particularly those involving permanent magnet machines, without using rare-earth materials. This can help reduce the use of strategic elements, and also can help achieving high temperature withstanding ability by these electromechanical devices. In fact, rare-earth free materials are important for some advanced battery systems, hence this area will help with battery technology.

Appendix

Appendix 1: Specifications of Komatsu Mining Truck

Information given below is taken from the Komatsu data sheet [3]. Although exact details on the systems are normally proprietary information of the manufacturers, some commonality can be noticed between these systems and some reasonable idea about the sizes of the components can be obtained, based on the information in these publicly available specification sheets.

AC current	
Alternator	**GTA-39**
Dual impeller in-line blower	453 m^3/min 16,000 cfm
Control	AC Torque Control System
Motorized wheels	GDY106 AC Induction Traction Motors
Ratio	32.62:1
Speed (maximum)	64.5 km/h 40 mph

Appendix 2: Accumulators

Accumulators basically consist of cylinders which have valves to let the gas in or out with appropriate safety mechanisms. Other types of accumulators are the bladder or diaphragm types as shown in Fig. 20a, b. Bladder-type accumulator consists of a flexible bladder submerged in an incompressible fluid. Inside the bladder there is a gas, which can be pressurized to compress or depressurized to expand. There are several working mechanisms for the accumulators, for example, in the hydropneumatic scheme the liquid pushes a piston, which then pushes a gas stored in the cylinder and compresses it, whereby energy is stored. Piston separates the liquid from the gas.

In the diaphragm accumulator a diaphragm B is pushed by some liquid from the bottom, which then changes the volume of the gas in the upper part of B, that is, between the diaphragm and the cylinder itself. In a bladder accumulator, the bladder B contains the gas which can be compressed. When the liquid enters the cylinder from the bottom, the bladder can expand or contract, thus energy can be stored in the gas when compressed and taken out from the gas when expanded. The energy is transferred through the fluid itself.

Bibliography

1. Ucarol H, Kaypmaz A, Tuncay RN, Tur O (2005) A performance comparison study among conventional, series hybrid and parallel hybrid vehicles. web article. http://www.emo.org.tr/ekler/c792a8279211dec_ek.pdf
2. Caterpillar. http://catsays.blogspot.com/2005/01/caterpillar-797b-mining-truck.html
3. Komatsu product brochure. http://www.komatsuamerica.com/default.aspx?p=equipment&f1=view&prdt_id=947
4. Wikimedia http://commons.wikimedia.org/wiki/File:Liebherr_T282.jpg (also see Liebherr product brochure, http://www.liebherr.com/ME/en-GB/products_me.wfw/id-14270-0
5. Komatsu product brochure. http://www.komatsu.com/CompanyInfo/press/2008051315113604588.html
6. Brown GN, Ebacher BJ, Koellner WG, Increased productivity with AC drives for mining excavators und haul trucks. http://info.industry.siemens.com/data/presse/docs/2050e.pdf
7. http://www.siemens.cz/extra/msv/cd/TS/Vozidla/Diesel-Electric_Loco_SD70MAC_EN.pdf
8. Kargul JJ (2007) Hydraulic hybrids – demonstration for port yard hostlers, in EPA presentation. www.epa.gov/otaq/technology, Accessed 11 July 2007
9. Bosch Rexroth product brochure, Bent axis hydraulic motor, Rexroth – Bosch Group, Axial Piston Fixed Displacement Motor AA2FM (A2FM), Bosch. https://www.boschrexroth.com/various/utilities/contact/index.jsp?cont_country=5&oid=3901
10. Olaer product brochure, Diaphragm and bladder accumulators, Olaer Inc., “Bladder and diaphragm accumulator product brochure, including High-pressure diaphragm accumulator conforming to EC regulations”, ELM Series. http://www.olaer.fr/en/recherche.php?idr=13, http://www.olaer.fr/en/recherche.php?idr=11
11. http://gas2.org/2009/09/18/caterpillar-builds-worlds-first-hybrid-bulldozer/
12. http://www.cleanmpg.com/forums/showthread.php?t=33855
13. Kwak S, Kim T, Park G (2010) Phase-redundant-based reliable direct AC/AC converter drive for series hybrid off-highway heavy electric vehicles. IEEE Trans Veh Technol 59(6):2674–2688
14. http://commons.wikimedia.org/wiki/File:14_stryker.jpg (Stryker)
15. http://commons.wikimedia.org/wiki/File:HEMTT.jpg (HEMTT)
16. http://commons.wikimedia.org/wiki/File:Two_M-3_Bradleys.jpg (Bradley)
17. http://commons.wikimedia.org/wiki/File:M1A1_Abrams_Tank_in_Camp_Fallujah.JPEG (Abrams)
18. http://commons.wikimedia.org/wiki/File:BundeswehrFennek.jpg (Fennek - European)
19. http://commons.wikimedia.org/wiki/File:FBI_Mine_Resistant_Ambush_vehicle.jpg (MRAP)
20. http://commons.wikimedia.org/wiki/File:Gladiator_RSTA.jpg (Gladiator)

Hybrid Energy Storage Systems for Vehicle Applications

MEHRDAD EHSANI
Department of Electrical & Computer Engineering,
Texas A&M University, College Station, TX, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Electric Load Characteristics
Candidates of Vehicle Energy Storage
Hybrid Energy Storage
Battery/Ultracapacitor System
Flywheel/Ultracapacitor System
Future Directions
Bibliography

Glossary

Electric load A device or system connected to the electrical power circuit for the purpose of demanding power.

Energy storage A device or system capable of storing energy in one of many physical forms.

Hybrid A combination of two or more items sharing a common function.

Hybrid energy storage A combination of two or more energy storage devices with complimentary capabilities.

Non-traction load Power demand for all purposes other than traction.

Traction load Power demand for the purpose of propelling the vehicle.

Definition of the Subject

The electric load of a vehicle can be decomposed into two components – static and dynamic load. The static component is slowly varying power with limited magnitude, whereas the dynamic load is fast varying power with large magnitude. The energy storage system, accordingly, comprises of two basic elements. One is energy source to support the static load and other is a power source to support the dynamic load. A smart combination of the available energy storages, which have different characteristics, may result in a high-performance energy storage system.

Introduction

It is widely agreed that vehicle electrification will lead to revolutionary improvements on vehicle performance, energy resource conservation, and pollution emissions. High-quality energy storage system is one of the most crucial components. Unfortunately, the currently available chemical, electrical, and mechanical energy storage systems, including various kinds of chemical batteries, ultracapacitors, and flywheels, cannot meet these requirements alone. In this entry, the possibility of composing a high-energy, high-power hybrid energy storage system is presented based on the analysis of inherent characteristics of different energy storage methods. The basic components in this system are chemical batteries, ultracapacitors, and flywheels (mechanical batteries). The objective is to fully utilize the advantages and suppress the disadvantages of each component to compose a compact energy storage system with high energy capacity and high power capability [3, 4, 6, 10].

Electric Load Characteristics

The electric load in a hybrid vehicle comprises of traction load and non-traction load. Regarding traction load, the energy storage is only responsible to supply a intermittent peak power which may be from a few seconds, such as in hard acceleration, steep hill climbing, obstacle negotiation, etc., to several minutes, such as in cross-country operation, medium hill climbing, etc. The vehicle in the former operation needs the energy storage to supply peak power in a few seconds with high magnitude, whereas in the latter situation, the vehicle needs the energy storage device to supply power in several minutes with medium magnitude. In normal road operation, the primary power source (engine or fuel cell) should be able to produce sufficient power to support the vehicle operation and, at the same time, produce additional power to charge the energy storage device to recoup its past discharge [7].

Similar to the traction motor load, the non-traction electric load can also be characterized by continuous power demand and intermittent peak power demand. Non-traction electric load is the total of electric loads that are used to serve the vehicle operation and auxiliaries. For a conventional vehicle, the non-traction electric loads consist of continuous loads, prolonged loads, and intermittent loads. The continuous loads, includes engine ignition, fuel injection, instrument panel, fuel pump, etc. These may need around 200 W. The prolonged loads, mainly including all the lights, may need about 260 W. Intermittent loads, mainly including, heater, front and rear wipers, brake light, electric windows, radiator cooling fan, heater blower motor, rear fog light, electric seat movement, electric mirror, etc., may need a total power of 1,700 W. The total of the non-traction electric peak power could reach about 2,140 W. However, the average is about 610 W. It is certain that more non-traction electric loads will be added in near future vehicles. They may include direct electrically driven power steering, direct electrically driven air conditioning, direct electrically engine cooling pump, electrically actuated braking, electrically driven active suspension, electrically driven engine valve actuator, etc. The total peaking power for the future electric loads may reach 20–25 kW while the average power may be only about 5 kW, one fourth or one fifth of the peak power [10].

The average of the electric loads needs the energy storage device to possess sufficient energy to support the vehicle's long-time operation, and the peak of the electric loads needs the energy storage device to be capable of delivering sufficient power in a short time period (usually less than 1 min).

Candidates of Vehicle Energy Storage

As mentioned above, the basic requirement for vehicle energy storage device is to have sufficient energy and also be able to deliver high power for a short time period. With the present technology, chemical batteries, ultracapacitors, and flywheel system (mechanical battery) are the main candidates for the vehicle energy storage device.

Chemical Batteries

The chemical battery is an energy storage device that stores energy in the chemical form and exchanges its energy with outside devices in electric form. Basically, a cell of battery has one positive electrode, which accepted electrons, one negative electrode, which gives out electrons and an electrolyte, which conducts ions and isolates electrons between the electrodes. The battery operation (discharging and charging) follows principles of electrochemistry and thermodynamics [2, 6, 9, 13].

Theoretical Specific Energy of Chemical Batteries

The amount of energy stored in a battery is completely determined by the amount of active reactants inside the battery, or more precisely, the free energy, ΔG, of the chemical reaction. The commonly used terms to measure the energy capacity of a battery are Watt-hours per kilogram (Wh/kg) called specific energy capacity and Watt-hours per liter (Wh/L) called energy density. The theoretical specific energy (maximum available) can be expressed as

$$E_{\text{specific}} = \frac{-\Delta G}{3.6 \sum M_i} \quad (\text{Wh/kg}), \tag{1}$$

where ΔG is the free energy of the reaction in a battery cell and ΣM_i is the sum of the molecular weight of the individual species involved in the battery reaction. The free energy, ΔG of a reaction is closely associated with the number of electrons transferred in the reaction and the electric potential (voltage), V_r, by

$$\Delta G = -n\,F\,V_r, \tag{2}$$

where n is the number of electrons transferred in the reaction and F is the Faraday constant, 96,495 ampere-seconds per gram-equivalent. Thus, the specific energy can be written as

$$E_{\text{specific}} = \frac{V_r\,F\,n}{3.6 \sum M_i} \quad (\text{Wh/kg}). \tag{3}$$

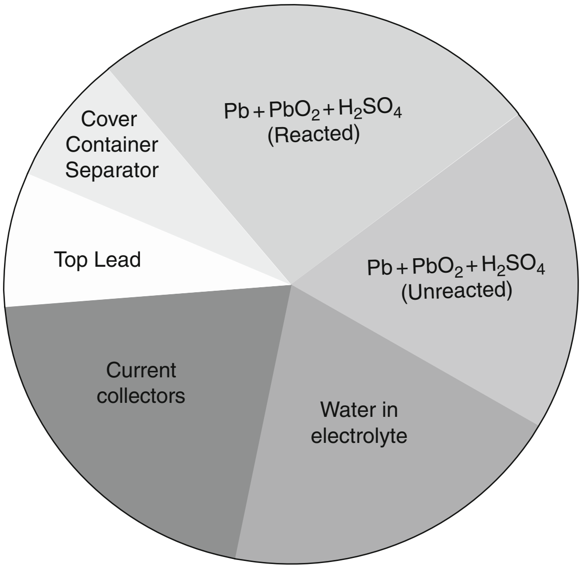
From theoretical electrochemistry and thermodynamics, the reaction potential of a battery cell can be described by

$$V_r = V_r^0 - \frac{RT}{nF} \ln \left[\frac{\Pi(\text{activities of products})}{\Pi(\text{activities of reactants})} \right], \tag{4}$$

where V_r^0 is a standard voltage at standard conditions when the battery is fully charged, R is gas constant, 8314 *J/°K. mole* and T is absolute temperature. The value in square brackets is nearly unity when the battery is charged, increasing with the process of battery discharging. When the reaction reaches equilibrium, V_r becomes zero, and the battery is fully discharged.

Practical Specific Energy of Chemical Batteries

Generally, practical batteries have much lower specific energies than their theoretical values. Figure 1 shows an example of the weight distribution of the components



Hybrid Energy Storage Systems for Vehicle Applications. Figure 1
Weight distribution of the components of a lead-acid EV battery with a specific energy of 45 Wh/kg at the $C_5/5$ rate [13]

of a lead-acid EV battery with a specific energy of 45 Wh/kg at the $C_5/5$ rate. In this example, only about 25% of the total weight of the battery is useful, that is, the practical specific energy is only about 25% of the theoretical one. Table 1 shows the data for some existing batteries.

Specific Power of Chemical Batteries

Specific power is the maximum power, per unit battery weight, that the battery can produce in a short period. The specific power of a chemical battery depends mostly on the battery internal resistance. With the battery model as shown in Fig. 2, the maximum power that the battery can supply to the load is

$$P_{\text{peak}} = \frac{V_0^2}{4(R_c + R_{\text{int}})}, \tag{5}$$

where R_c is the conductor resistance (Ohmic resistance) and R_{int} is the internal resistance caused by chemical reaction.

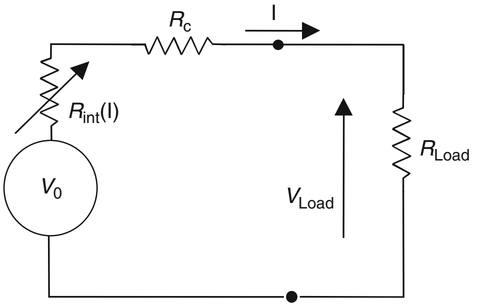
Internal resistance R_{int} represents the voltage drop, ΔV, which is associated with the battery current. The voltage drop ΔV, termed overpotential in battery terminology, includes two components: one is caused by reaction activity ΔV_A and the other caused by electrolyte concentration ΔV_C. General expressions of ΔV_A and ΔV_C are

$$\Delta V_A = a + b \log I \tag{6}$$

Hybrid Energy Storage Systems for Vehicle Applications. Table 1 Status of battery systems for EV

	Specific energy (Wh/kg)	Peak power (W/kg)	Energy effi.(%)	Cycle life	Cost ($/kW.h)
Acidic aqueous solution					
Lead-acid	35–50	150–400	>80	500–1,000	120–150
Alkaline aqueous solution			75	800	250–350
Nickel/cadmium	40–60	80–150	65	1,500–2,000	200–400
Nickel/iron	50–60	80–150	70	300	100–300
Nickel/zinc	55–75	170–260	70	750–1,200+	200–350
Ni-metal-hydride	70–90	200–300	<50	?	?
Aluminum/air	200–300	90	60	500+	50
Iron/air	80–120	90	60	600+	90–120
Zinc/air	100–200	30–80			
Flow					
Zinc/bromine	70–85	90–110	65–75	500–2,000	200–250
Vanadium redox	20–30	110	75–85	-	400–450
Molten salt					
Sodium/Sulfur	150–240	230	85	800+	250–450
Sodium/Ni/Cl	90–120	130–160	80	1,200+	230–345
Li-ion-sulfur (FeS)	100–130	150–250	80	1,000+	110
Organic/lithium					
Lithium-ion	80–130	200–300	>95	1,000+	200
Li-ion high power[a]	85–95	~4,000	–	–	–
Li-ion high energy[a]	135–150	~600	–	–	–

[a] Reported by SAFT [14]



Hybrid Energy Storage Systems for Vehicle Applications. Figure 2
Battery circuit model

and

$$\Delta V_C = -\frac{RT}{nF}\ln\left(1-\frac{I}{I_L}\right), \tag{7}$$

where a and b are constants, R is gas constant, 8314 $J/°K.\ mole$, T is the absolute temperature, n is the number of electrons transferred in the reaction, F is the Faraday constant, 96,495 ampere-seconds per gram-equivalent, and I_L is the limit current. Accurate determination of battery resistance by analysis is difficult. Specific powers are usually obtained by measurement.

Table 1 shows the status of battery systems potentially available for EV [13]. It can be seen that although specific energies are high in advanced batteries, the specific powers show no significant improvement. About 300 W/kg might be the optimistic estimate. Recent Li-ion high-power and high-energy batteries have been reported with as high as 4,000 and 150 Wh/kg, respectively. However, their practicality and cost need much improvement.

Energy Loss and Thermal Management

The presence of internal resistance in batteries causes energy loss in the battery discharging and charging process. This dissipated energy causes the battery temperature to rise, shortening the battery life. The energy loss can be expressed as

$$E_{\mathrm{loss}} = R_{\mathrm{batt}} I^2, \tag{8}$$

where $R_{\mathrm{batt}} = R_c + R_{\mathrm{int}}$ is the battery internal resistance and I is the battery current. The battery resistance is usually related to the battery temperature and battery state-of-charge. Low temperature usually leads to high battery resistance because of less reaction activities of the reactants. Thermal management is usually needed to maintain the battery in an appropriate temperature range. Discharging at low state-of-charge and charging at high state-of-charge can also lead to high battery resistance because of the nonuniformity in the concentration of the reactants. Thus, the battery should be managed to operate in a middle range of state-of-charge, such as 40–60%. Furthermore, high charging and discharging current should be avoided so that the energy loss in battery can be reduced (refer to Eq. 8), and consequently, the thermal management can be simplified.

Ultracapacitors

The ultracapacitor is another candidate for energy storage in the hybrid vehicle. The ultracapacitor is characterized by high specific power, high efficiency, excellent temperature adaptability, and long service life. However, it suffers from very limited specific energy. The characteristics of a 42-V ultracapacitor pack specifically designed for HEV are listed in Table 2 [11]. Reference [8] has reportedly developed 2700-F ultracapacitors with 2.5 V of cell voltage, 2.55 kW/kg of specific power, and 3.23 Wh/kg of specific energy. It is believed that the unit power cost of the ultracapacitor is lower than that of batteries. However, due to its low specific energy and its high dependence of the state-of-charge on voltage, it is difficult to use ultracapacitors alone as the energy storage on hybrid vehicles. The most promising approach is to hybridize the ultracapacitors with other energy storage devices, such as batteries or flywheels as will be discussed in following sections.

Flywheel

Flywheel is an energy storage device which directly stores mechanical energy in an extra high-speed rotating wheel. Because of the difficulty of using the mechanical energy of the flywheel directly for traction, a high-speed electric machine is usually connected to the flywheel to constitute a system, called "mechanical battery," as shown in Fig. 3. The flywheel can be charged and discharged through the electric machine. The major advantages of flywheel system are their

Hybrid Energy Storage Systems for Vehicle Applications. Table 2 Characteristic data of an ultracapacitor

Rated capacitance (DCC[a], 25°C)	[F]	145
Capacitance tolerance	[%]	±20
Rated voltage	[V]	42
Surge voltage	[V]	50
Max. series resistance, ESR (DCC, 25°C)	[mΩ]	10
Specific power density (42 V)	[W/kg]	2,900
Max. current	[A]	600
Max. stored energy	[J]	128,000
Specific energy density (42 V)	[Wh/kg]	2.3
Max. leakage current (12 h, 25°C)	[mA]	30
Weight	[kg]	15
Volume	[L]	22
Operating temperature	[°C]	−35 to +65
Storage temperature	[°C]	−35 to +65
Life time (25°C)	[y]	10 ΔC <20% of initial value, ESR <200% of initial value
Cyclability (25°C, I = 20 A)		500,000, ΔC <20% of initial value, ESR <200% of initial value

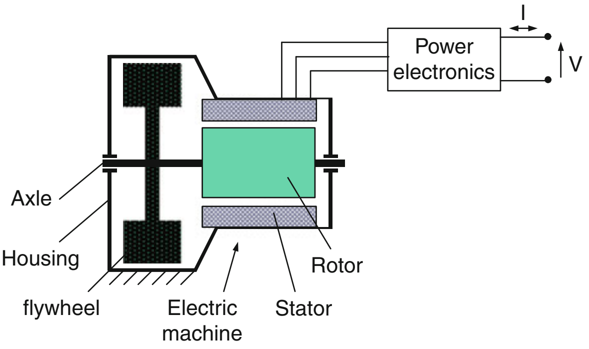
[a] *DCC* discharging at constant current

Source: Ref. [11]

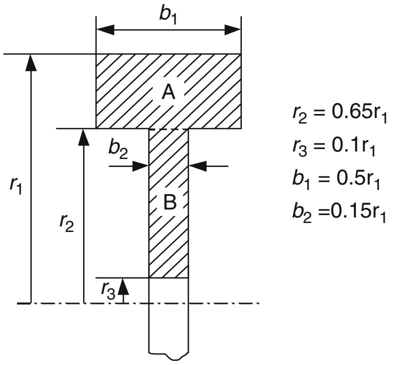
relative high energy capacity and low environmental dependence, especially temperature. However, high speeds require high-strength materials, special design bearings (mostly magnetic bearings), and high vacuum housing to reduce the aerodynamic drag. These requirements may cause technical difficulties in their application to vehicles.

Energy Capacity of Flywheel System

The energy stored in a high-speed flywheel is expressed as



Hybrid Energy Storage Systems for Vehicle Applications. Figure 3
Configuration of a flywheel system (mechanical battery)



Hybrid Energy Storage Systems for Vehicle Applications. Figure 4
Geometry of a flywheel

$$E_f = \frac{1}{2} J_f \omega_f^2, \tag{9}$$

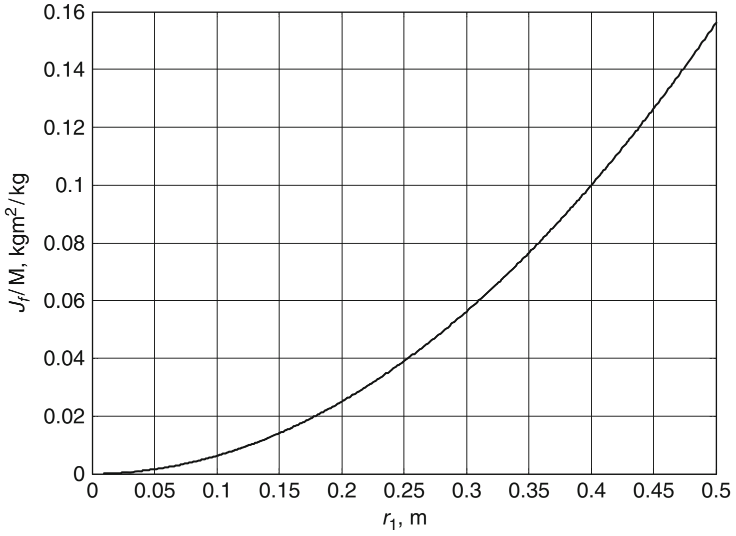
where J_f is the moment of inertia of the flywheel in kgm^2 and ω_f is the angular velocity of the flywheel in *rad/s*. High energy capacity demands high speed (usually 50,000–60,000 rpm) and high moment of inertia from the flywheel.

The moment of inertia of a flywheel per unit flywheel mass (weight) depends exclusively on its geometric design (refer to Fig. 4), as expressed by

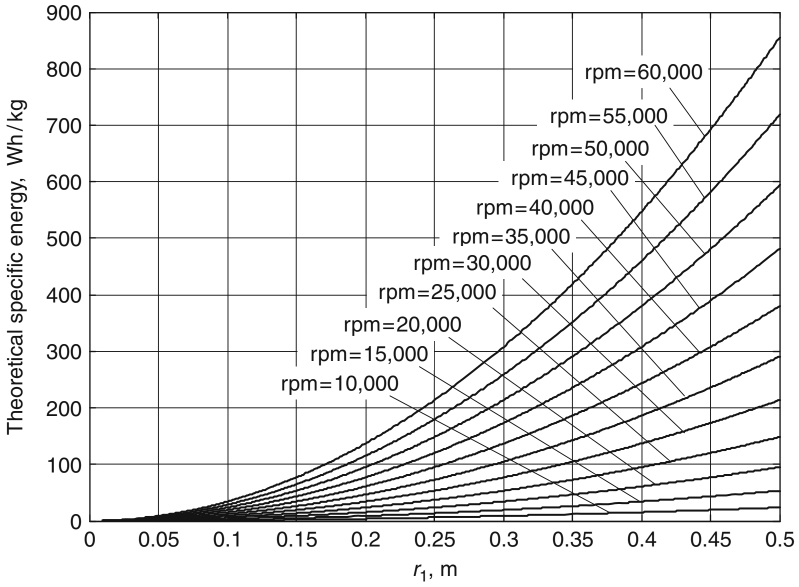
$$\frac{J_f}{M_f} = \frac{1}{2}\left[\frac{b_1(r_1^4 - r_2^4) + b_2(r_2^4 - r_3^4)}{b_1(r_1^2 - r_2^2) + b_2(r_2^2 - r_3^2)}\right]. \tag{10}$$

The effective approach to increase the J_f/M_f is to increase b_1 and r_1. That is, to put as much mass as far away from its rotating center as possible. This is usually limited by the rotor material strength. So high-strength materials are usually employed in high-speed flywheels. Figure 5 shows the spatial size of a flywheel, which is represented by r_1, versus the moment of inertia of the flywheel per unit mass, using the geometric numbers in Fig. 4.

The theoretical specific energy of a flywheel can be calculated by Eqs. 9 and 10 for the geometrically specified flywheel in Fig. 4. Figure 6 shows the theoret-



Hybrid Energy Storage Systems for Vehicle Applications. Figure 5
Outer radius of the flywheel versus the moment inertia of the flywheel per unit mass



Hybrid Energy Storage Systems for Vehicle Applications. Figure 6
Theoretical specific energy of flywheel versus flywheel geometry size, r_1

ical specific energy of the flywheel versus the flywheel geometric size r_1 and flywheel angular speed. This figure indicates that with increasing flywheel size, r_1, and angular speed, the specific energy can be increased significantly. For a midsize vehicle, $r_1 = 035$ m and $rpm = 50000$ may be good estimates. In this case, the theoretical specific energy of flywheel will reach about 300 W/kg.

In practice, the effective mass used to store energy in a flywheel system only takes part of the total system mass. The ineffective masses, such as the housing, the stator of the electric machine, the bearings, etc., and the less effective mass, such as the rotor of the electric machine, may make up to over half of the total system mass. Thus, the practical specific energy of a flywheel system is much lower than its theoretical value and may be in the range of 100–150 Wh/kg, which is in the same range as that of advanced chemical batteries, such as NiMH and lithium batteries, but much high than that of the traditional batteries, such as lead-acid and nickel/cadmium batteries.

Power Capability of Flywheel System

The power of a flywheel system can be expressed as

$$P_f = \frac{dE_f}{dt} = J_f \omega_f \frac{d\omega_f}{dt}, \tag{11}$$

where $T_m = J_f \frac{d\omega_f}{dt}$ is the torque of the electric machine to accelerate (charging) or decelerate the flywheel (discharging). Thus, the power capability of the flywheel system is completely determined by the power capability of its electric machine.

The electric machine used in the flywheel system should be have high speed, high efficiency, light weight, and small volume. Further, in order to simplify the control of the vehicle electric loads, such as traction motors, actuators, etc., the terminal voltage of the flywheel system should be maintained approximately constant in the intended operating speed range of the flywheel. Electric machine theory indicates that varying the machine speed while keeping a constant terminal voltage can be accomplished by controlling its magnetic field. Thus, wide speed operating range with constant voltage requires the machine to have wide "field weakening" capability and a long constant power speed range.

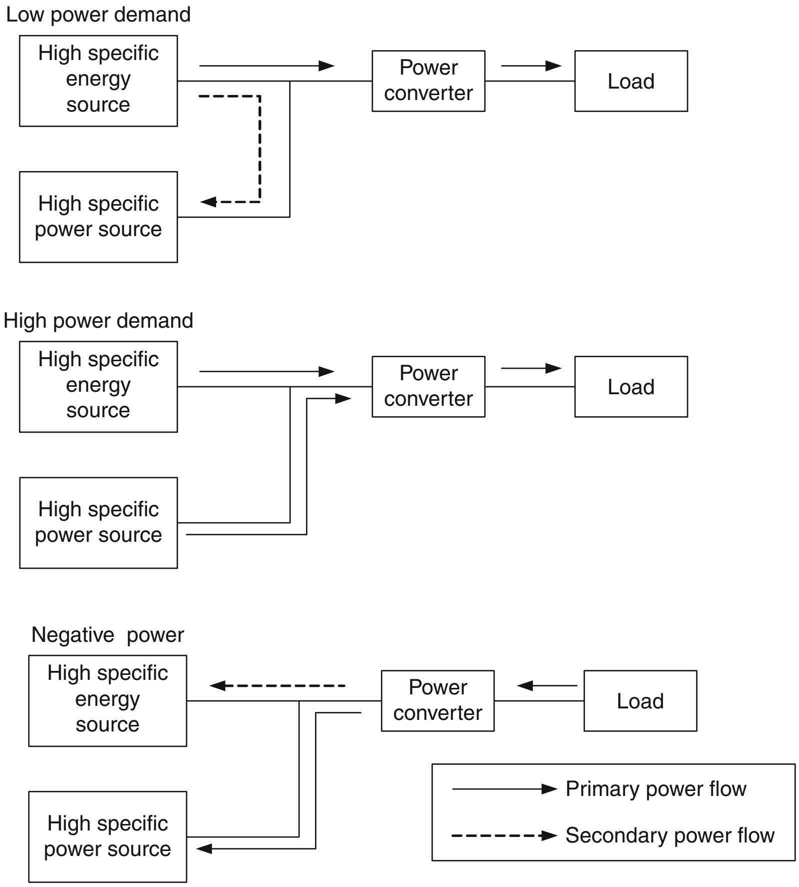
Referring to the motors that are appropriate for electric and hybrid electric vehicles, the induction motor (IM), permanent magnet (PM), brushless dc motor (BLDC), and switched reluctance motor (SRM) seem to be possible candidates for application to the flywheel system. The induction motor is the most mature technology among these motors. Its maximum speed can be extended up 3–4 times of its base speed by using field oriented control technology [1]. The BLDC is well known for its lightweight, small volume, and high efficiency due to its permanent magnet field. However, the extent of its speed range beyond its base speed is very limited due to its limited capability of the field weakening, resulting in a limited constant voltage speed range. The SRM is well known for its simple structure, straightforward control, excellent reliability, high speed, and long maximum speed beyond its base speed. These inherent properties make the SRM an attractive candidate for application to the flywheels system, notwithstanding its relative low torque density, relative to the BLDC. Based on our analysis and development of traction motors for EV and HEV [12], at Texas A&M University, it is believed that the specific power of the overall flywheel system will not be much more than 500 W/kg. This is similar to that of advanced lead-acid batteries and less than that of advanced high-power batteries, such as lithium batteries and ultracapacitors.

Efficiency of Flywheel System

Energy losses in the flywheel system are not mainly mechanical, such as friction and aerodynamic drag, but also electrical losses in its electric machine. With high vacuum inside the housing of the flywheel system and with magnetic bearings, it is believed that mechanical efficiency may reach 98–99%. The electric efficiency mostly depends on the electric motor/generator, whose efficiency may be around 85–90%. Thus, the total efficiency of the flywheel system may be around about 83–87%, which is similar to that of advanced high-efficiency batteries and higher than that of conventional batteries, but much less than that of ultracapacitors.

Hybrid Energy Storage

As discussed above, chemical batteries, ultracapacitors, and flywheels each have their specific operating


Low power demand
High specific energy source
High specific power source
Power converter
Load
High power demand
High specific energy source
High specific power source
Power converter
Load
Negative power
High specific energy source
High specific power source
Power converter
Load
Primary power flow
Secondary power flow

Hybrid Energy Storage Systems for Vehicle Applications. Figure 7
Conceptual operation of the hybrid energy storage

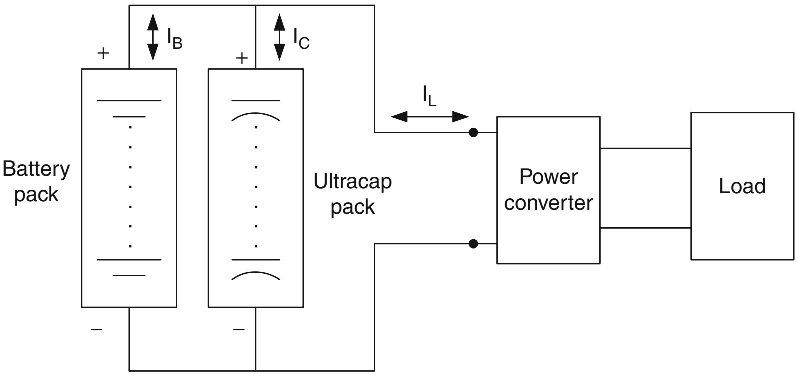
characteristics. The chemical battery and flywheel are more for high energy storage and deliver their energy with medium power for a relatively long period of time. Whereas, the ultracapacitor is more for high-power source and delivers its high power in a relatively short period of time. It is possible, by combining two of these basic energy storage technologies together to constitute a high-energy and high-power storage system. This is the hybrid energy storage system as is conceptually shown in Fig. 7.

Battery/Ultracapacitor System

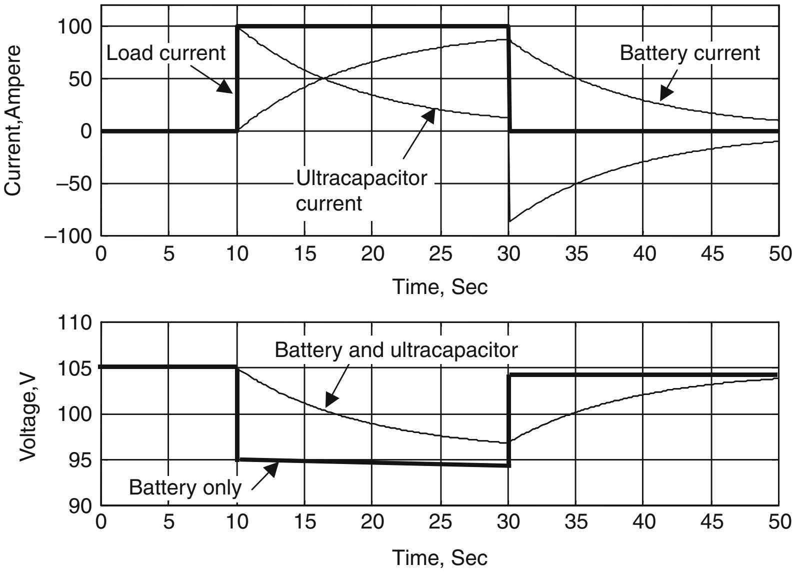
Passive Interconnection

The most straightforward approach is to directly connect ultracapacitors to batteries as shown in Fig. 8. This configuration has the simplest structure and no control unit. The behavior of the ultracapacitors is more like a current filter. Thus, the high-peak battery currents are leveled, as illustrated in Figs. 9 and 10. In Fig. 9, a step load current is applied to the storage system. First, this step current is in the ultracapacitors and then the load current gradually transfers from the ultracapacitors to the batteries due to the voltage drop in the ultracapacitors. When the load current disappears, the ultracapacitors are charged back up by the batteries automatically. Figure 10 shows the simulation results of the battery current and the ultracapacitor current while a random load current applied.

The leveled battery current, due to the ultracapacitors, can have many benefits for the batteries. These include smaller battery pack, higher battery



Hybrid Energy Storage Systems for Vehicle Applications. Figure 8
Passively connected batteries/ultracapacitors system



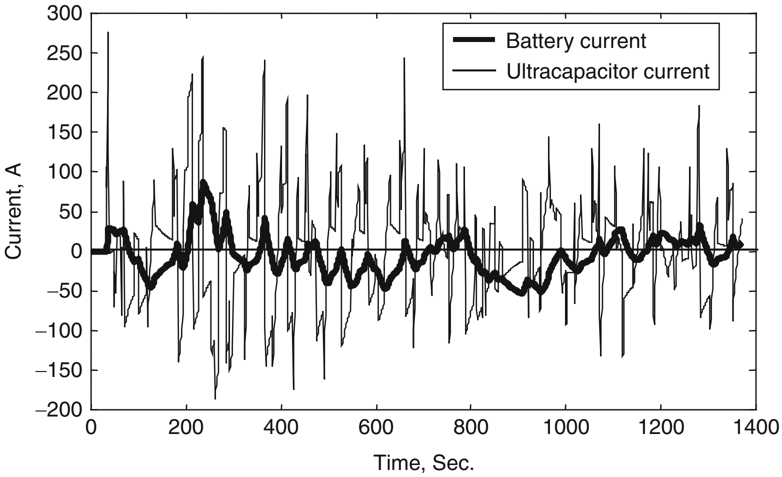
Hybrid Energy Storage Systems for Vehicle Applications. Figure 9
Battery current and ultracapacitor current with a step load current input

operating efficiency, easier battery thermal management, and longer battery life. The limited variation of hybridized battery terminal voltage can also ease the control of the connected load power converter. With proper design of all the components, this simple hybrid structure may yield a high-power, high-energy, and high-efficiency energy storage system.

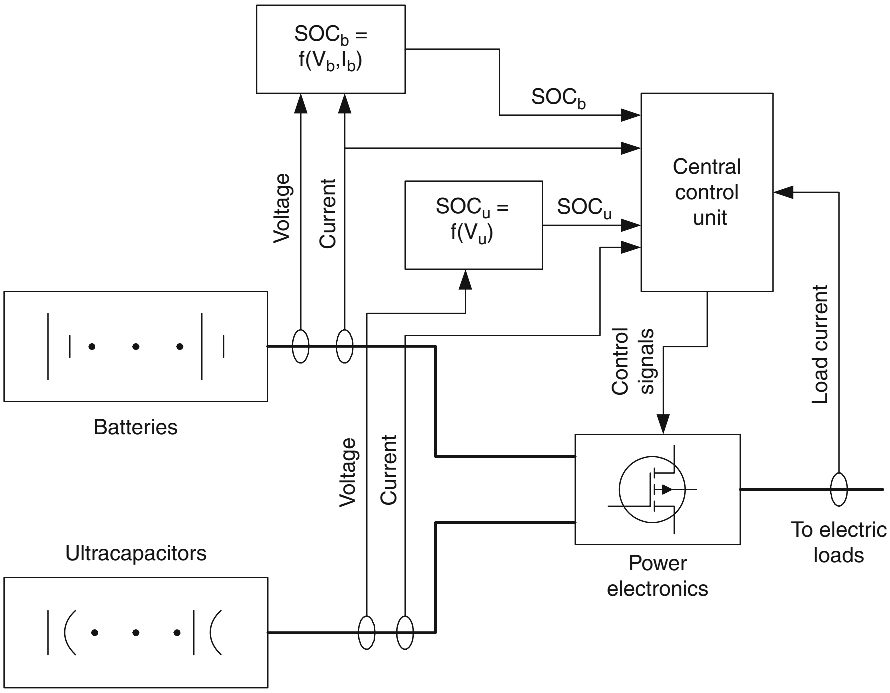
Active Interconnection

Clearly, in the passively connected battery/ultracapacitor system, the power flow between the battery pack, the ultracapacitor pack, and the load cannot be managed. Consequently, the high power benefits of the ultracapacitors may not be best utilized. Figure 11 shows the concept of an actively controlled battery/ultracapacitor system. In this system, a power electronic regulator is used to manage the power flow between the battery pack, the ultracapacitor pack, and the load.

Basically, the power conditioning operation can be divided into three different modes: (1) ultracapacitor peaking operation for high power demand, (2) ultracapacitor charging from batteries, and (3) battery alone operation. These operation modes are



Hybrid Energy Storage Systems for Vehicle Applications. Figure 10
Battery current leveled by passively connected ultracapacitors



Hybrid Energy Storage Systems for Vehicle Applications. Figure 11
Conceptual illustration of the actively controlled battery/ultracapacitor system

implemented by a central control unit. The central control unit commands the power electronics, according to the control strategy and the received signals from the current and voltages sensors, as shown in Fig. 11. The control objectives are (1) to meet the load power requirement, (2) to keep the battery current in a preset band, and (3) to keep the battery state-of-charge, SOC, in its safe region (e.g., 0.4–0.6), where the battery efficiency are usually optimized.

This system can fully use the high power property of the ultracapacitors, resulting in a smaller battery pack. Further, the actively controlled battery current can lead to more efficient battery operation and easier thermal management.

Flywheel/Ultracapacitor System

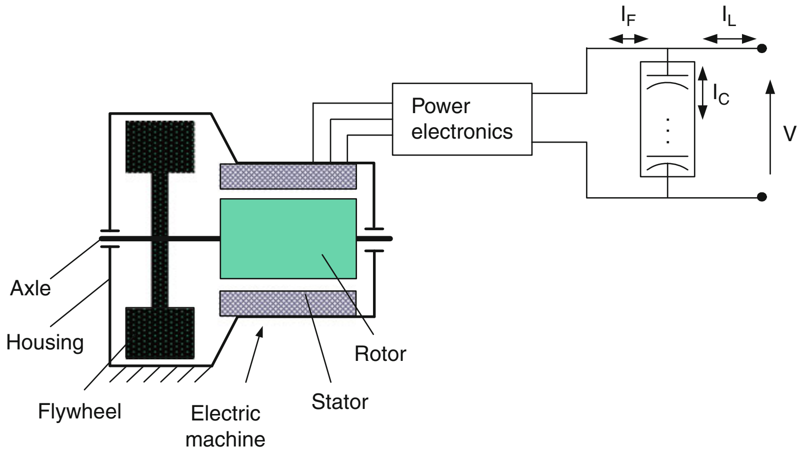
As discussed in the flywheel section, the flywheel system, similar to batteries, can be categorized as an energy intensive storage device. This is mainly due to its power capability being constrained by its electric machine. Integration of the high-power ultracapacitors with the high-energy flywheel system may result in a high-power, high-energy and high-efficiency storage system.

As before, the ultracapacitors can be either passively connected to the flywheel system, as shown in Fig. 12, or through an electronic power management unit, as shown in Fig. 13. The passively connected flywheel/ultracapacitor system may be expected to have a performance similar to that of the passively connected battery/ultracapacitor system, as shown in Figs. 9 and 10. The actively controlled flywheel/ultracapacitor system can be expected to have the advantage of a smaller flywheel electric machine, stable operation, and high efficiency.

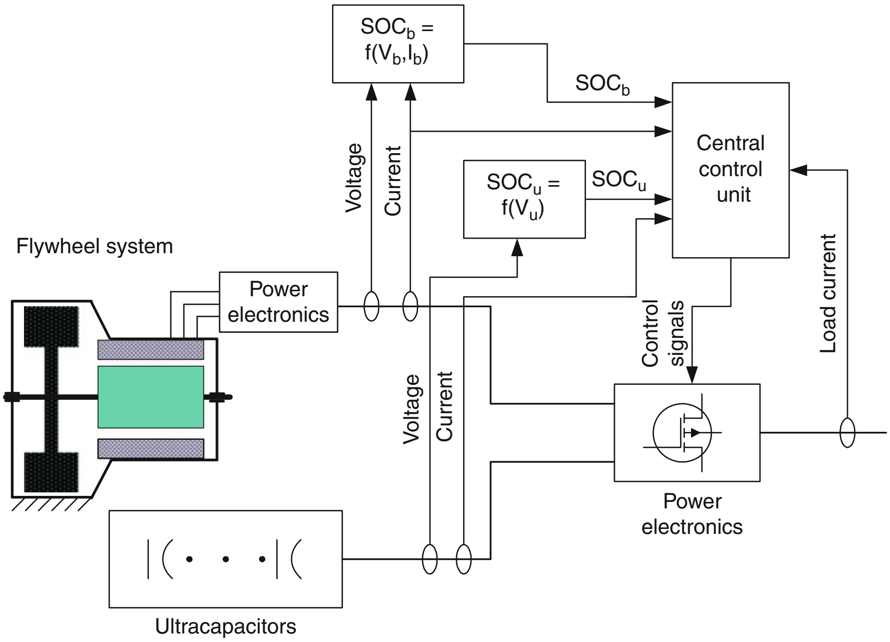
Compared with battery/ultracapacitor hybrid system, flywheel/ultracapacitor system would have the additional advantage of extended service life, wide temperature tolerance, and lower maintenance. These advantages are very important in vehicle applications.

Future Directions

The vehicle energy storage should be able to supply sufficient energy and power to meet both the steady and dynamic load requirements. Thus, high specific energy and high specific power are necessary to achieve compact vehicle energy storage. Chemical batteries and mechanical flywheels can be categorized as energy sources and ultracapacitors as power sources. By hybridizing these storage devices, a high-energy and high-power-energy storage system may be developed. In the passively connected battery/ultracapacitor hybrid system, the battery peak currents can be significantly moderated. Consequently, the battery efficiency, service life, and thermal management can be significantly improved. In the actively controlled battery/ultracapacitor hybrid system, the power flow between the batteries, ultracapacitor and loads can be actively controlled. Therefore, the operation of the whole system can be better optimized. Flywheel/ultracapacitor hybrid systems, either passively connected or actively controlled, can have the advantages of compact design, high efficiency, extended service life, wide temperature tolerance, and low maintenance requirements. These advantages are very important for advanced vehicles.



Hybrid Energy Storage Systems for Vehicle Applications. Figure 12
Passively connected flywheel/ultracapacitor system



Hybrid Energy Storage Systems for Vehicle Applications. Figure 13
Actively controlled flywheel/ultracapacitor system

Although all of the above systems have been used individually in EV and HEV designs, the hybridized concepts, described here remain to be applied to the commercial EV and HEV products of the future. As the ultracapacitor technology and applications become more mature, their use in vehicle-hybrid-energy storage applications is likely to spread. However, as the power capability of chemical batteries continues to improve, the use of hybrid energy storage becomes less imperative. The eventual outcome of the above competing developments is not clear at this time. However, hybrid energy storage may find uses in advanced vehicles in one form or another for traction and other loads.

Bibliography

1. Chan CC, Chau KT (2001) Modern electric vehicle technology. Oxford University Press, New York
2. Crompton TR (1996) Battery reference book, 2nd edn. Society of Automotive Engineering
3. Delphi Automotive Company, http://www.delphi.com/news/pressReleases/pr11607-04112002
4. Ehsani M, Gao Y, Butler KL (1999) Application of electrically peaking hybrid (ELPH) propulsion system to a full size passenger car with simulated design verification. IEEE Trans Veh Technol 40:6
5. Ehsani M et al (2003) Investigation of high-energy and high-power storage systems for military vehicle applications. SAE-FTT 2003
6. Gao Y, Ehsani M (2002) An investigation of battery technologies for the army's hybrid vehicle application. In: Vehicular technology conference, 2002. Proceedings VTC 2002-Fall. 2002 IEEE 56th, vol 3, pp 1505–1509
7. Kajs J (2001) Combat vehicle mobility requirements. Seminar at Texas A&M University
8. Maxwell Technologies, http://www.maxwell.com
9. Messerle HK (1969) Energy conversion statics. Academic, New York
10. Miller JM, Emadi A, Rajarathnam AV, Ehsani M (1999) Current status and future trends in more electric car power systems. In: Proceedings of the IEEE vehicular technology conference Houston
11. Montena, Montena components. http://www.montena.com
12. Rahman Z, Butler KL, Ehsani M (2000) Effect of extended-speed, constant-power operation of electric drives on the design and performance of EV-HEV propulsion system. SAE future car congress, Apr 2000, Crystal City, Paper No. 2000-01-01557
13. Rand DAJ, Woods R, Dell RM (1998) Batteries for electric vehicles. Society of Automotive Engineering
14. SAFT Batteries, http://www.saftbatteries.com/automotive/uk/f/f.htm
15. ThermoAnalytics, http://www.thermoanalytics.com/support/publications/batterytypesdoc.html

Hydrogen from Biomass

RALF SCHMERSAHL[1,2], MARCO KLEMM[3],
RUTH BRUNSTERMANN[4], RENATUS WIDMANN[4]
[1]Deutsches BiomasseForschungsZentrum GmbH, German Biomass Research Centre (DBFZ), Leipzig, Germany
[2]Institute of Agricultural Engineering Bornim e.V. (ATB), Potsdam, Germany
[3]German Biomass Research Centre (DBFZ), Leipzig, Germany
[4]Department of Urban Water and Waste Management, University of Duisburg-Essen, Essen, Germany

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Biological Hydrogen Production
Hydrogen Production by Solid Biomass Gasification
Future Directions
Bibliography

Glossary

Biohydrogen Hydrogen derived from biomass.

Biological biohydrogen generation Production of hydrogen based on a biological resp. biochemical conversion of biomass.

Biomass From a scientific and technical point of view, biomass is defined as material of biological origin excluding material embedded in geological formations and/or transformed to fossil.

Thermochemical biohydrogen generation Production of hydrogen based on a heat-induced resp. thermochemical conversion of biomass.

Definition of the Subject and Its Importance

Today, there is no disputing that the use of renewable energy has to be increased in order to reduce anthropogenic CO_2 emission as well as the dependence on the fossil fuels. Approximately 95% of the hydrogen produced today comes from carbonaceous raw material, but primarily with a fossil resource as original energy source. Only a fraction of this hydrogen is currently used for energy purposes; the bulk serves as feedstock for manifold purposes, e.g., in the petrochemical industry as well as for food, electronics, and metallurgical processing. However, the share of hydrogen in the energy market is increasing, and hydrogen production will need to keep pace with this growing market [1]. In this sense, this paper summarizes the state of the art of the most important processes, techniques, and research activities in the field of hydrogen production using biomass resources.

Introduction

Hydrogen (H_2) is a secondary energy carrier that has to be produced from other energy carriers. A more and more discussed option for H_2 production is the use of biomass. It may be carried out either by biological or thermochemical processes. The characteristics of thermochemical H_2 production differ significantly from those of biological processes. The differences cover the feedstock basis as well as the reaction chemistry, the plant design, and the plant scales.

In the case of biological H_2 production, H_2 may be produced as a photosynthesis product (e.g., by microalgae). A second option is the use of fermentative microbiological degradation pathways, where H_2 occurs as metabolism product. Both pathways are still in the research scale.

Thermochemical biomass processes, such as combustion, pyrolysis, and gasification, are applied to produce different forms of energy carriers and chemicals. From those basic processes, gasification complies best with the requirements for the production of H_2. However, none of the present biomass gasifier designs is a genuine H_2 supplier: the product gas needs sophisticated posttreatment techniques to achieve both high H_2 yield and a sufficient H_2 purity.

Within the following explanations, these two hydrogen provision routes are discussed in detail.

Biological Hydrogen Production

Biological hydrogen production can be carried out by a wide variety of different organisms [2]. The biological processes differ according to the essential boundary conditions. Below, the process principles as well as some examples are discussed.

Hydrogen from Biomass. Table 1 Possible biochemical reactions of biological hydrogen production processes [3]

Process	Biochemical reaction	Boundary conditions
Direct photolysis	$2H_2O + \text{light} \rightarrow 2H_2 + O_2$	E.g., green algae in pure cultures, sterile conditions, lack of sulfur inhibition of the enzyme Fe-hydrogenase by oxygen
Indirect photolysis	$12H_2O + 6CO_2 + \text{light} \rightarrow C_6H_{12}O_6 + 6O_2$	Photoautotrophic bacteria, e. g., Cyanobacteria in pure culture, sterile conditions, enzymes: hydrogenase
	$C_6H_{12}O_6 + 2H_2O \rightarrow 4H_2 + 2CH_3COOH + 2CO_2$	
	$2CH_3COOH + 4H_2O + \text{light} \rightarrow 8H_2 + 4CO_2$	
Dark fermentation	Biomass + $H_2O \rightarrow H_2 + CO_2$ + VFA (e.g., acetic acid)	Exclusion of methane-producing bacteria, mixed culture generated from different origin
Photofermentation	$CH_3COOH + 2H_2O + \text{light} \rightarrow 4H_2 + 2CO_2$	Photoheterotrophic bacteria (e.g., purple bacteria), enzyme nitrogenase, lack of nitrogen

Process Principles

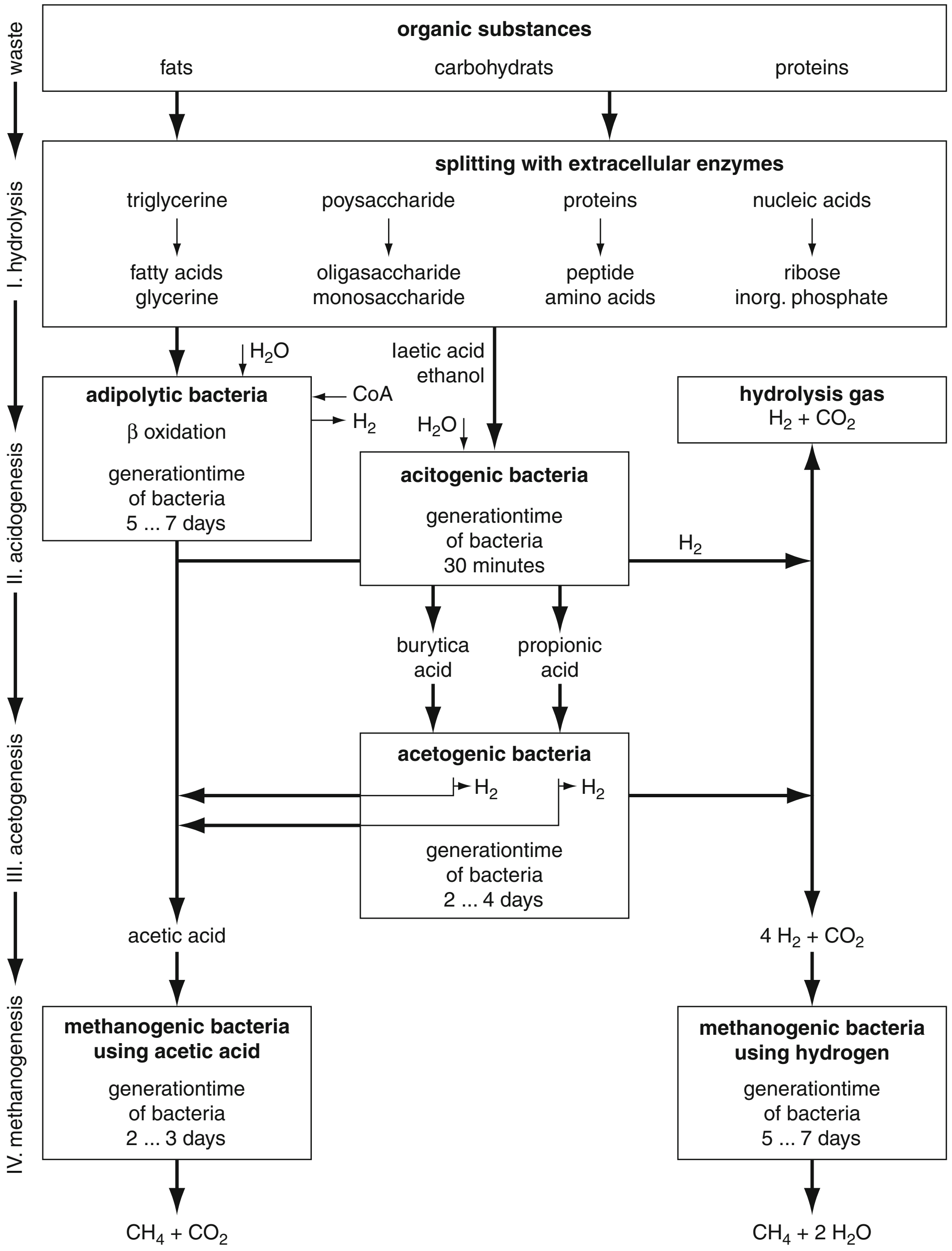
An overview about different biohydrogen production methods is given in Table 1. Biohydrogen production by photolysis and dark fermentation and photofermentation are described and discussed below.

Photolysis Hydrogen production by the microalgae (i.e., *Chlamydomonas reinhardtii*) has been studied intensively in the recent years. This algae use the direct photolysis pathway to produce hydrogen out of water, if the microorganisms are stressed by lack of sulfur. Under these conditions, the algae produce hydrogen and oxygen but cannot grow. By this reason, a second stage with optimal conditions for the growth of the algae is necessary. Therefore, carbon dioxide is needed to build organic compounds like carbohydrates by photosynthesis. Additionally, by this pathway, hydrogen and oxygen are produced simultaneously, and the produced oxygen inhibits the enzyme hydrogenase, which is involved in the process. The color of the photosystems of the algae and phototrophic bacteria limits the dimension of the reactor because the penetration of light is reduced by the microorganisms itself [12]. For this reason, different reactor types are designed, for example, tubular or flat panels with a diameter or depth respectively of 10 cm at maximum [5]. Most of the experiments are done in lab scale by using artificial light. An advantage of this process is the clean energy production, because only hydrogen and oxygen are produced by this process.

However, the development of photobiological algae systems for biohydrogen production is still fundamental research. Due to the fact that the produced hydrogen yield per lightened area is very low – 1.5% based on a full solar spectrum [4] – further investigations are necessary to develop economic, viable, large-scale applications. Via genetic modification and cloning of the microorganisms, higher yields may be achieved [5]. Apart from that, H_2 and O_2 are produced simultaneously, which requires careful handling.

In contrast to the direct photolysis, it is possible during indirect photolysis to separate the oxygen formation by photosynthesis from hydrogen production in either time or space; this process is divided in three steps (Table 1). First, the photosynthesis takes place. In this step, CO_2 is needed, and oxygen is produced. Biochemical reaction of dark fermentation and photofermentation takes place in the second and third steps. This avoids the inhibition of the enzyme by oxygen. Still, the overall efficiency of the process is low like direct photolysis [4].

Dark Fermentation Dark fermentation is based on anaerobic degradation pathway, which is currently used for methane production with different biomass as feedstock. The biological process is divided in four steps. Depending on the composition of the biomass, different degradation pathways are used, and different microorganisms are involved. In steps 2 and 3, hydrogen is produced as an intermediate in several degradation pathways by a variety of microbes. In a conventional anaerobic digestion, this hydrogen is transformed into methane by methanogenic bacteria



Hydrogen from Biomass. Figure 1
Degradation pathway during anaerobic digestion

in the fourth step. The degradation pathways during anaerobic digestion depending on the composition of feedstock are shown schematically in Fig. 1. There, the biomass is divided in fats, carbohydrates, and proteins because the anaerobic degradation pathway depends on the structure of molecules.

If the hydrogen consumption is stopped by elimination of the hydrogen-consuming microorganisms (i.e., the methanogenic bacteria), hydrolysis gas accrue, which consist mainly of hydrogen and carbon dioxide. The methanogenic microorganisms might be eliminated by heat treatment of the start culture, for example, digested sludge [6, 7], or by using a seed material without strict anaerobic hydrogen-consuming bacterial-like compost or soil [3, 8] or pure stains of hydrogen-producing bacteria [4].

An opportunity to prohibit the hydrogen consumption by using a mixed culture out of digested sludge is to optimize the applied condition for hydrogen production and suppress hydrogen consumption. This is possible because methane is generated at a pH value between 6.0 and 8.0, whereas the hydrolytic and acidogenic bacteria prefer pH value between 4.5 and 6.0 for the degradation of carbohydrates. Nevertheless, the generation time of the methane-producing bacteria is within 2–7 days, much higher than the generation time of hydrogen-producing bacteria. Due to the fact that the generation time of carbohydrate-degrading acidogenic bacteria is only 30 min at optimal conditions, the hydraulic retention time (HRT) might be shorter. In contrast to the hydrogen-producing bacteria, the methane bacteria cannot accumulate in the reactor at a hydraulic retention time less than 48 h. These facts indicate as well that carbohydrates are most suitable for the dark fermentation process by using mixed cultures, which are generated out of digested sludge.

The process might be driven in the mesophilic [3, 7], thermophilic [6], or hyperthermophilic temperature range. Every temperature has its own advantages and disadvantages.

The extreme thermophilic anaerobic bacteria provide the highest hydrogen yields, reaching almost a maximum of 4.0 mol H_2/mol of glucose by using pure stains. Using pure cultures implicates that the feedstock has to be sterilized to avoid an accumulation of other microorganisms in the reactor. Additionally, the energy demand of the process is higher than at lower temperatures. The above-mentioned hydrogen yield of 4.0 mol H_2/mol glucose (this corresponds with 498.38 Nl H_2/kg glucose) is only reached if the combination of feed (different kinds of sugar) and microorganisms is optimal [4, 10, 11]. That implicates that a small change of feedstock (e.g., hexose instead of glucose) results in less efficiency of the plant. For this reason, a pretreatment step for the conversion of more complex biomass to simple sugars is necessary, which reduce the efficiency of the plant as well.

Hence, with the diversity in a mixed culture of strict anaerobic bacteria like Clostridia and facultative anaerobic hydrogen-producing bacteria, a variety of biomass can be converted to hydrogen. This process may be realized in the mesophilic or thermophilic temperature range. Due to the less energy demand, better stability of the process and approximately the same hydrogen yields [3], the mesophilic temperature range seems to be most promising for a technical application with a mixed bacteria culture as seed material and biowaste as feedstock in short term.

Beside the hydrogen yield (given in Nl hydrogen per kg of substrate), the gas and especially the hydrogen production rate is important for a technical application. Due to the fact that the bacteria density has to be low as well as the organic concentration in the influent, only the reduction of hydraulic retention times (HRT) provides an increase of hydrogen production rate. Hence, different reactor types were used. Continuous stirred-tank reactor (CSTR) is the easiest technique, which is similar to the used technique in wet anaerobic digestion. But the hydraulic retention has to be higher than the generation time of the microorganisms to provide an accumulation of these bacteria. Therefore, the hydraulic retention time has to be higher than 12 h.

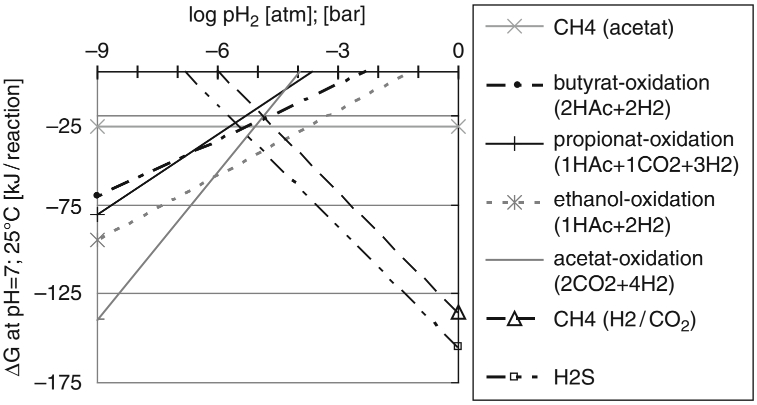
A disadvantage of dark fermentation is the inhibition by its end product hydrogen. This is caused by biochemical equilibrium and thermodynamics of reactions. Figure 2 shows the energy yield of different reactions based on the hydrogen partial pressure, which affects the solubility of hydrogen in the liquids. Only if the hydrogen partial pressure is less than 0.1 bar will the conversion of organic acids to hydrogen become exothermic.

To provide these conditions, the produced hydrogen has to be removed from the system continuously. This can be done by:

- Purging with nitrogen or another inert gas
- Stirring with a high frequency

The purging with inert gas results in an increase of gas production. In parallel, the hydrogen concentration within the produced gas is decreasing, and therefore, an upgrading before the use (e.g., in a fuel cell) is necessary.

In general, hydrogen production by dark fermentation is restricted by the incomplete degradation of organic matter into volatile fatty acids, hydrogen, and carbon dioxide. Consequently, a further conversion of the liquid residues into an energy carrier is necessary.



Hydrogen from Biomass. Figure 2
Energy yield of reactions dependent on hydrogen partial pressure

Photofermentation Photofermentation is driven by photoheterotrophic microorganisms, like purple bacteria (i.e., different strains of *Rhodobacter*). These bacteria are unable to split water. However, under anaerobic condition, these bacteria are able to degrade simple organic components like volatile fatty acids and alcohols into hydrogen and carbon dioxide under nitrogen-limited circumstances. The catalyst of the process is nitrogenase, which is sensitive to oxygen like hydrogenase. There are certainly no problems because no oxygen is produced. Normally, this enzyme catalyzed the fixation of molecular nitrogen (N_2) to ammonia NH_3, which is needed by the bacteria to grow. In absence of nitrogen, the enzyme is able to catalyze the evolution of hydrogen. The produced gas consists of hydrogen (80–90% (v/v)) and carbon dioxide (10–20% (v/v)) [4, 5].

The process depends on light. So, the reactor needs a high surface-to-volume ratio like the direct and indirect photolysis. By using sunlight, the increase of temperature has to be taken into account to prevent the system from damage by high temperature.

The above-mentioned limitation to convert only organic acids and alcohols by using of light as an energy source makes it a possible second treatment step after dark fermentation.

Comparison So far, all biohydrogen production technologies have been tested in lab scale. The feasibility of a technical implementation for hydrogen-producing photosystems is currently restricted by space requirements [12]. Additionally, the surface-to-volume ratio is restricted by the penetration of light into the reactor. Table 2 shows a comparison of required dimensions of bioreactors to produce enough gas to power a PEM fuel cell.

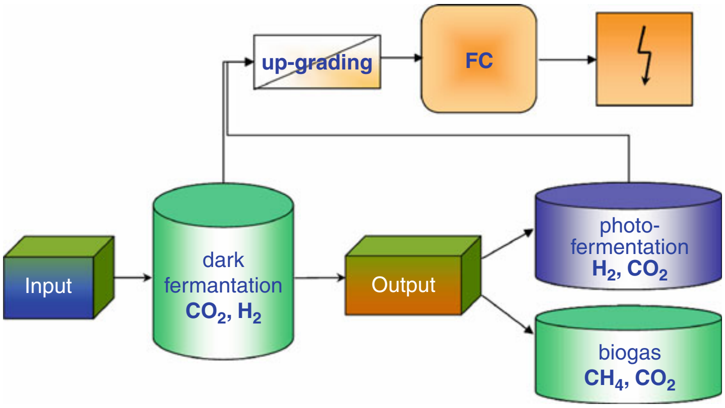
Up to now, very few technical and economic studies for biological hydrogen production systems have been carried out [4, 5, 13].

Regarding space requirements, no algae photosystem provides a sufficient amount of hydrogen in relation to the necessary reactor size. Without intensive further research and development of the methods, a technical realization in the large scale seems to be impossible within the next decade. Additionally, most of the photosystems work with pure strains of microorganisms and lack of nutrients (sulfur or nitrogen). Therefore, antiseptic techniques are needed to avoid an accumulation of other bacteria or yeast. Furthermore, circumstances in respect to nutrient limitations have to be taken into account.

Although the process of biohydrogen production by dark fermentation has been investigated very intensively during the last decade, there are still a lot of open questions. Most of the research has been carried out in batch test systems, just small continuous systems have been operated yet. The reactor dimensions usually do not exceed a working volume of a few liters. Additionally, the dark fermentation is an incomplete degradation of organic matter with volatile fatty acids in the liquid phase. Consequently, the dark and photofermentation are promising processes. In case of dark fermentation, a second treatment step is necessary. In case of photofermentation, a pretreatment step is required.

Hydrogen from Biomass. Table 2 Size of bioreactor required to power PEM fuel cells 0

System	Size of bioreactor (m^3) required to power a PEM fuel cell of			
	1.0 kW	1.5 kW	2.5 kW	5.0 kW
Direct photolysis	341	512	856	1,710
Indirect photolysis	67.3	101	169	337
Photofermentation	149	224	374	758
Dark fermentation min.	0.198	0.297	0.495	0.989
Dark fermentation max.	2.91	4.38	7.31	14.62


up-grading
FC
Input
dark fermantation CO_2, H_2
Output
photo-fermentation H_2, CO_2
biogas CH_4, CO_2

Hydrogen from Biomass. Figure 3
Biological-based biomass to hydrogen concepts

Biomass to Hydrogen Systems

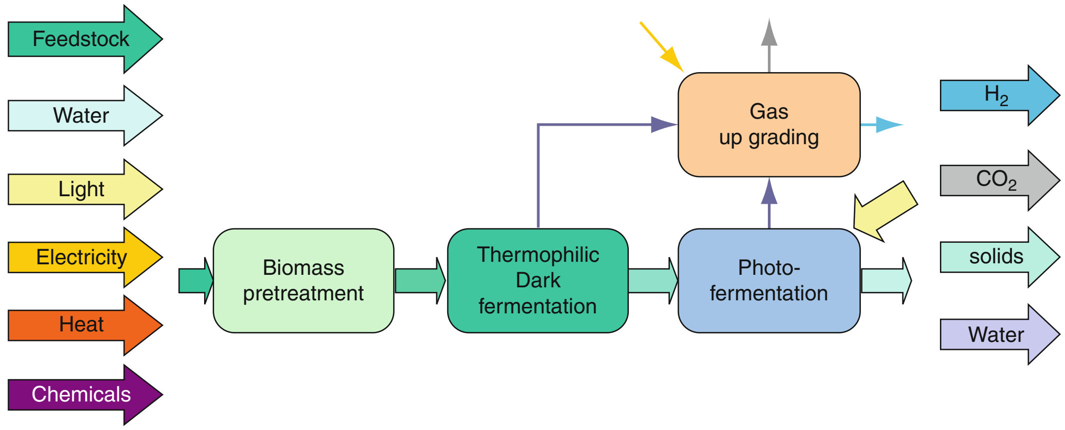
A complete biological conversion into hydrogen and carbon can only be realized by a combined multistage treatment (e.g., dark fermentation as the first stage and conversion of the produced volatile fatty acids into hydrogen by photofermentative microorganisms as the second stage [4]). For implementation, even combined dark fermentation/methane fermentation plants or a material utilization of the reactor effluent is conceivable.

With conventional anaerobic digestion, an energy carrier in form of biomethane is already established. Therefore, methane fermentation combined with steam reforming will lead to energy losses by the hydrogen production step. Hence, it should be carefully considered whether methane could be used directly for energy generation, e.g., in a suitable fuel cell (MCFC, SOFC) or in the mobility sector (gas upgrading to natural gas quality).

Below, some examples of combined systems are discussed in detail (Fig. 3).

Combination of Dark Fermentation and Photofermentation The combined dark fermentation and photofermentation process connects two new technologies with high research requirements because they have only been realized in lab scale and not in pilot or technical scale yet. This will cause additional problems for technical implementation [9]. Figure 4 shows a scheme of the combined dark and photofermentation system.

Caused by the specific microorganisms and to increase the conversion of biomass to hydrogen during dark fermentation with pure bacteria strains, a pretreatment step is used. Here, hydrolytic bacteria and/or addition of chemicals should degrade a variety of biomass (especially carbohydrates) to simple molecules. The thermophilic temperature range is supplied



Hydrogen from Biomass. Figure 4
Scheme of combined dark and photofermentation [9] (simplified)

Hydrogen from Biomass. Table 3 Biochemical reactions for the combination of dark and photofermentation with glucose as substrate

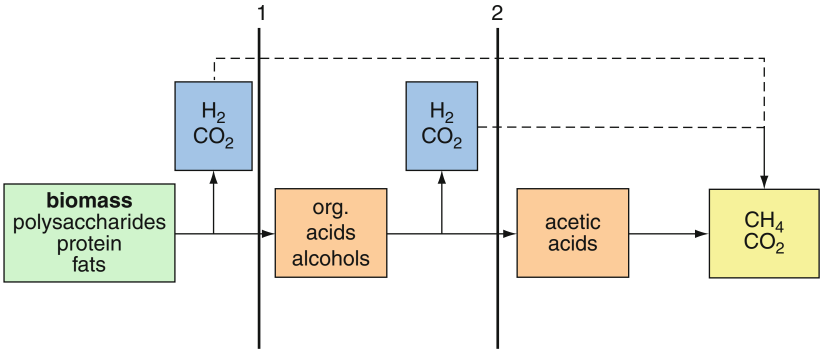
Substrate	First reactor (dark fermentation)	Second reactor (photofermentation)
$C_6H_{12}O_6$ + 2 H_2O→	$2CH_3COOH$ + **$2CO_2 + 4H_2$** [3, 4]	**8 H_2 + 4 CO_2** [14]
	$2CH_3COOH$ + $4H_2O$ + Light→	

by heating the system. This process operates with specific anaerobic bacteria specialized to degrade only different kinds of sugar or starch to H_2, CO_2, and organic acids. The thermophilic temperature in this stage provides sanitation and decreases the competition with other microorganisms. Additionally, chemicals are used in this stage for nutrient supply and pH stabilization. The effluent contains mainly volatile fatty acids and alcohols dissolved in water. This effluent is pumped in photobioreactor. Here, the light-dependent photofermentation takes place. The biochemical reactions in this combined system, except the biomass pretreatment, are shown in Table 3.

Combined Dark and Methane Fermentation Hydrogen yields of 4 mol/mol glucose (Eq. 1) at maximum are possible by dark fermentation with only 2 mol acetates as by-product [2, 15]. If only butyric acid occurs as by-product, the hydrogen yield decreases to 2 mol/mol glucose (Eq. 2) [15]. Typically, the end product of dark fermentation by using mixed microflora is a mix of different volatile fatty acids like acetic and butyric acid [7, 16]. Accordingly, the biochemical possible hydrogen yield is 2.5 mol/mol glucose (Eq. 3).

If the liquid effluent of the H_2 reactor is used as carbon source in a methane reactor, the organic acids become degraded to CO_2 and CH_4. The theoretical methane yield that can be generated depends on the liquid products of the hydrogen stage. If the acidogenesis takes place in the first reactor, only acetic acids would be transferred into the methane reactor, and only these acids would be converted into CH_4 and CO_2. This case is signed by line 2 in Fig. 5. Line 1 indicates that only the hydrolysis and acetogenesis take place in the first reactor. If so, the hydrogen yield decreases, and the methane yield increases according to Eq. 2 and Eq. 3. The stoichiometries of some possible biochemical reactions for the combined hydrogen and methane production with glucose as carbon source are given in Table 4. The chemical equation for the optimal degradation of glucose by a conventional fermentation (Eq. 4) is also given in this table.

An engineered approach towards a technical system has not been made yet. However, up to now, the combined hydrogen/methane fermentation system



Hydrogen from Biomass. Figure 5
Combined biological hydrogen and methane production out of biomass

Hydrogen from Biomass. Table 4 Biochemical equation for the combined biological hydrogen and methane production with glucose as energy source

	Substrate	First reactor (H_2)	Second reactor (CH_4)
Eq. 1	$C_6H_{12}O_6 + 2\ H_2O \rightarrow$	$2CH_3COOH$ **+ $2CO_2$ + $4H_2$** [2, 15]	
		$2CH_3COOH \rightarrow$	**$2CH_4 + 2CO_2$**
Eq. 2	$2\ C_6H_{12}O_6 \rightarrow$	$2C_3H_7COOH$ **+ $4CO_2$ + $4H_2$** [17]	
		$2C_3H_7COOH + 2H_2O \rightarrow$	**$5CH_4 + 3CO_2$**
Eq. 3	$8\ C_6H_{12}O_6 + 4\ H_2O \rightarrow$	$4CH_3COOH + 6C_3H_7COOH$ + **$16CO_2$+ $20H_2$** [7, 16]	
		$4CH_3COOH + 6C_3H_7COOH + 6H_2O \rightarrow$	**$19CH_4 + 13CO_2$**
Eq. 4	$C_6H_{12}O_6 \rightarrow$		**$3CH_4 + 3CO_2$**

offers the highest realization potential regarding the biological hydrogen production methods. This is due to the fact that only one pretreatment module has to be implemented in a proved technology.

Hydrogen Production by Solid Biomass Gasification

Thermochemical hydrogen production can be carried out by a wide variety of different processes. Below, the basic principles as well as some examples are discussed.

Process Principles

Generally, gasification is a thermochemical process that converts a carbon-containing substance, here lignocellulosic biomass, by partial oxidation into a fuel gas. A gasification agent (i.e., air, oxygen, or steam) provides the required oxygen. While air gasification generates a low-calorific-value gas suitable for power generation, oxygen and steam gasification lead to a medium-calorific-value gas, referred to as syngas. Syngas is suitable for the production of energy carrier such as H_2, methane, and liquid hydrocarbons.

The overall gasification reaction for biomass may be expressed after Eq. 5.

$$\begin{aligned}\text{Biomass} + H_2O + O_2 &\rightarrow H_2 + CO + CO_2 \\ &+ CH_4 + \text{hydrocarbons} + \text{char}\end{aligned} \tag{5}$$

The heat of reaction (ΔH_R) is positive for steam gasification (endothermic reaction) and negative for oxygen gasification (exothermic reaction). Reaction stoichiometry favors steam as gasification agent for a maximum H_2 output, but pure steam gasification requires an external heat supply adding up to a concept called allothermal steam gasification. At first glance, oxygen gasification is comparatively less attractive for H_2 production since less H (in form of steam) enters the gasifier. But other aspects have to be considered that prohibit such a simple conclusion:

- Oxygen gasification achieves higher temperatures enabling higher carbon-to-gas ratios.
- Autothermal operation is advantageous for overall process efficiency since heat transfer losses are smaller.
- Steam addition in downstream processing may balance the initial lower H content.

Biomass gasification typically operates under thermodynamic conditions that imply an almost complete carbon conversion into gaseous species (i.e., CO, CO_2, and CH_4). Assuming carbon conversion is complete, the equilibrium gas composition is given by two homogenous gas reactions, methane steam reforming (Eq. 6) and water–gas shift (Eq. 7).

$$CH_4 + H_2O \leftrightarrow CO + 3H_2 \quad \Delta H_R = +206\ \text{KJ/mol} \tag{6}$$

$$CO + H_2O \leftrightarrow CO_2 + H_2 \quad \Delta H_R = -41\ \text{KJ/mol} \tag{7}$$

In practice, carbon conversion is restricted due to mass transfer limitations and reaction kinetics. The raw syngas often contains light and heavy hydrocarbons being intermediates in terms of volatilization of the biomass organics. Especially aromatic compounds, known as tars, require a sophisticated conversion process or in depth removal before downstream gas processing.

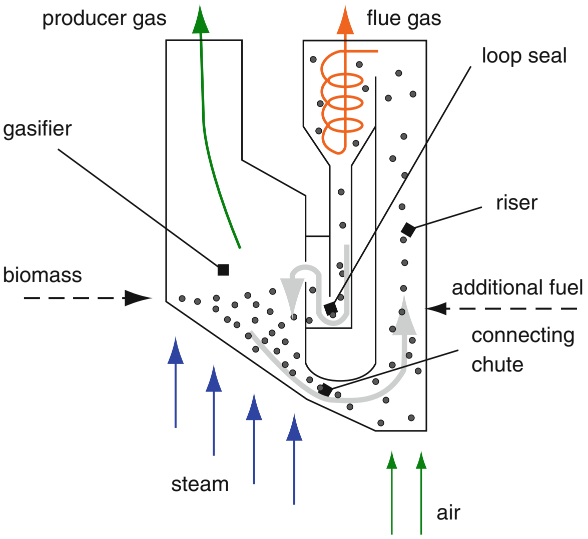
Biomass gasification for fuel production relies on different gasification concepts, namely entrained flow, fluidized bed, and dual fluidized bed steam gasification:

- Fluidized bed gasification is realized by a bed of fine, inert or catalytic active, material that is fluidized by blowing oxygen upwards through the bed. In contrast to power and heat applications, air is an improper gasification agent as nitrogen dilutes the syngas, boosting the downstream gas treatment effort. Depending on the gas velocity, the bed material is agitated (bubbling fluidized bed, BFB) or suspended throughout the gasifier (circulating fluidized bed, CFB). The syngas leaves the reactor at the top. In case of CFB, it has to be separated from the bed material with a cyclone. Fluidized bed gasifiers allow for pressurized operation. Typical gasification temperatures are in the range of 850–950 C. Compared to other gasification options, the main advantages of CFB gasifiers are high cold gas efficiencies, low investment costs, and a wide applicable power range of some tens to some hundreds of megawatts. Major drawbacks are related to comparatively low H_2 content in syngas and limited operating pressure because of feeding and fluidization issues.
- Entrained flow gasification takes on a special position in biomass gasification. In contrast to fluidized bed gasifier designs, it refers to established fossil fuel technologies for large-scale implementation. The key measure to enable entrained flow gasification for biomass is to manipulate the biomass properties by pyrolysis processes, such as torrefaction and carbonization, in a way that they become more similar to coal. The resulting bio-oil, char, or slurry are pressurized and then atomized with pure oxygen in a burner at the top of the gasifier. Gasification temperatures are in a range of 1,300–1,600°C, high enough to convert ash into slag. Entrained flow biomass gasification has some advantages compared to other gasification processes: (1) the high syngas pressure is favorable for H_2 purification and distribution, (2) the raw gas is tar-free and therefore requires only modest gas cleaning efforts, and (3) the possibility to combine decentralized, small-scale biomass pretreatment with centralized, large-scale gasification avoids the long-distance transport of raw biomass with a low energy density while still profiting from economy of scales. On the other hand, there are some major drawbacks: (1) the overall concept implies an efficiency penalty due to losses during thermochemical pretreatment, (2) formation and discharge of slag is difficult to handle since biomass feedstock provides a wide range of ash properties, and (3) the H_2 content in syngas is low, providing a heavy duty for downstream gas conditioning.
- Dual fluidized bed steam gasification provides a major advantage for H_2 production by increasing the amount of H entering the process. But because the gasification process is endothermic, an external

heat source and a heat transfer medium are required. Therefore, steam gasification concepts usually work with two fluidized beds, one for the gasification process and a second one for combustion. As an example, the process principle of the Fast Internally Circulating Fluidized Bed (FICFB) technology is shown in Fig. 6 0. Here, biomass is fed into the gasification chamber where it is converted into nitrogen-free syngas and char using steam as gasification agent. Char, recycled syngas, and recycled tars are burnt with air in the combustion chamber, heating the accompanying catalytic bed particles. The hot bed material is then fed back into the gasification chamber, providing the required reaction heat. Dual fluidized bed gasifiers operate at ambient pressure with a gasification temperature of 800–870°C [18].

In case of H_2 production, all presented biomass gasifiers require further downstream gas processing to achieve better H_2 yield und purity. Gas conditioning by steam reforming (Eq. 6) and water–gas shift (Eq. 7) utilize steam to generate additional H_2:

- Steam reforming is the process where methane and other hydrocarbons are converted into carbon dioxide, carbon monoxide, and H_2 in the presence of steam. In technical applications, reforming is performed utilizing a catalyst, which is usually nickel-based. Steam reforming should be performed at high temperatures to shift the equilibrium to the right side of Eq. 6. There should be an excess of steam to prevent coking. At temperatures below 900°C, most of the nickel catalysts are sensitive to sulfur poisoning. Methane steam reforming is strongly endothermic. As for the gasification process itself, the question arises, whether the heat is provided autothermal by oxygen supply or allothermal from an external source. Low pressure is favorable since it shifts the reaction equilibrium towards H_2.
- In contrast, the water–gas shift reaction is exothermic. Unfortunately, the excess heat is not suitable to support steam reforming because the temperature level is too low. Water–gas shift is often performed in two steps: a high temperature shift and a low temperature shift. The high temperature shift usually is performed at 300–400°C using iron–chrome catalysts, and the low temperature shift operates below 250°C utilizing copper and zinc oxide catalysts. The amount of carbon monoxide can be lowered to less than 1%. The equilibrium is not sensitive to pressure.



Hydrogen from Biomass. Figure 6
Fast Internally circulating fluidized bed (FICFB) gasification 0

Biomass gasification with downstream gas conditioning results in an H_2-rich gas that typically contains a significant amount of accompanying gas species (e.g., carbon dioxide, carbon monoxide, nitrogen, and a variety of trace components). Physical or chemical scrubbers for acid gas removal (AGR) permit a comparatively cheap and efficient bulk CO_2 removal, but high purity H_2 output requires (additional) pressure swing adsorption, membrane separation, or preferential CO oxidation upstream of the acid gas removal.

Pressure swing adsorption works on the principle of molecular sieving. Molecular sieves are zeolites or activated carbon with a defined pore structure allowing for the separation of molecules according to size and polarity. The gas is fed to a system of packed beds. Frequent pressurization of the beds leads to, alternately, adsorption and release of the unwanted gas species. After desorption, the bed is purged. In the time, the first bed is regenerated; the second bed is pressurized, after which the gas feed is switched back to the first bed.

Biomass to Hydrogen Systems

The basic concept of biomass gasification plants for H_2 production is presented in Fig. 7. It involves several main process steps: (1) biomass pretreatment, mainly drying and sizing, but in case of entrained flow gasification also thermochemical treatment such as fast pyrolysis or torrefaction; (2) the gasification process itself, performed either allothermal with a dual fluidized bed design or autothermal with fluidized bed or entrained flow gasifiers; and (3) downstream gas processing that covers gas cleaning for the removal of impurities, gas conditioning by reforming and water–gas shift reaction, and H_2 purification.

A biomass gasification plant is not energy autonomous. Most of the processes require a significant amount of electrical energy. Steam generation is also an issue. Depending on infrastructural and business conditions, it could be a favorable option to integrate power and heat generation. If so, some of the syngas is used to fire an engine or a gas turbine for the simultaneous generation of electricity, heat, and fuel. Another option is the conversion of surplus heat with a steam turbine. In the future, the operation of a fuel cell feed with H_2 is another possibility. If power and heat generation exceeds the demand of the plant, surpluses can be sold in addition to the H_2 product. This approach is referred to as polygeneration concept [19].

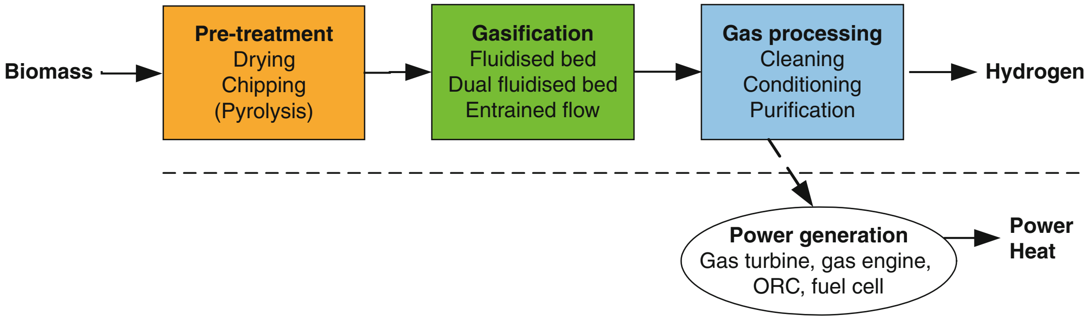
Present biomass gasification plants are originally designed for the production of biofuels other than H_2. But a proper downstream process design allows for H_2 production as well. If designed for H_2 output, the fuel synthesis island has to be replaced by H_2 purification processes.

In case of entrained flow gasification concepts, biomass pretreatment and the gasification process itself will be the same for all biofuel output option. Gasification parameters do not allow for further optimization of the H_2 yield because there is a trade-off between efficiency and H_2 yield: a higher steam input requires more oxygen to achieve the necessary operating temperatures. Carbon conversion in entrained flow gasifiers is almost complete. The hot raw gas consists of H_2, CO, and CO_2. It is cooled by a water quench using the sensible heat of the gas stream for water vaporization. The amount of water fed into the gas stream lowers the demand for additional steam during the water–gas shift.

H_2 production asks for an as high as possible CO conversion since the energy content of unconverted CO is lost for H_2 output. Therefore, a staged water–gas shift layout is essential. Available catalysts are sensitive to sulfur poisoning so that sulfur components have to be removed upstream. Depending on the requested H_2 quality, different H_2 purification processes have to be applied. Figure 8 provides at the left side a basic scheme of a process chain for H_2 production based on entrained flow gasification.

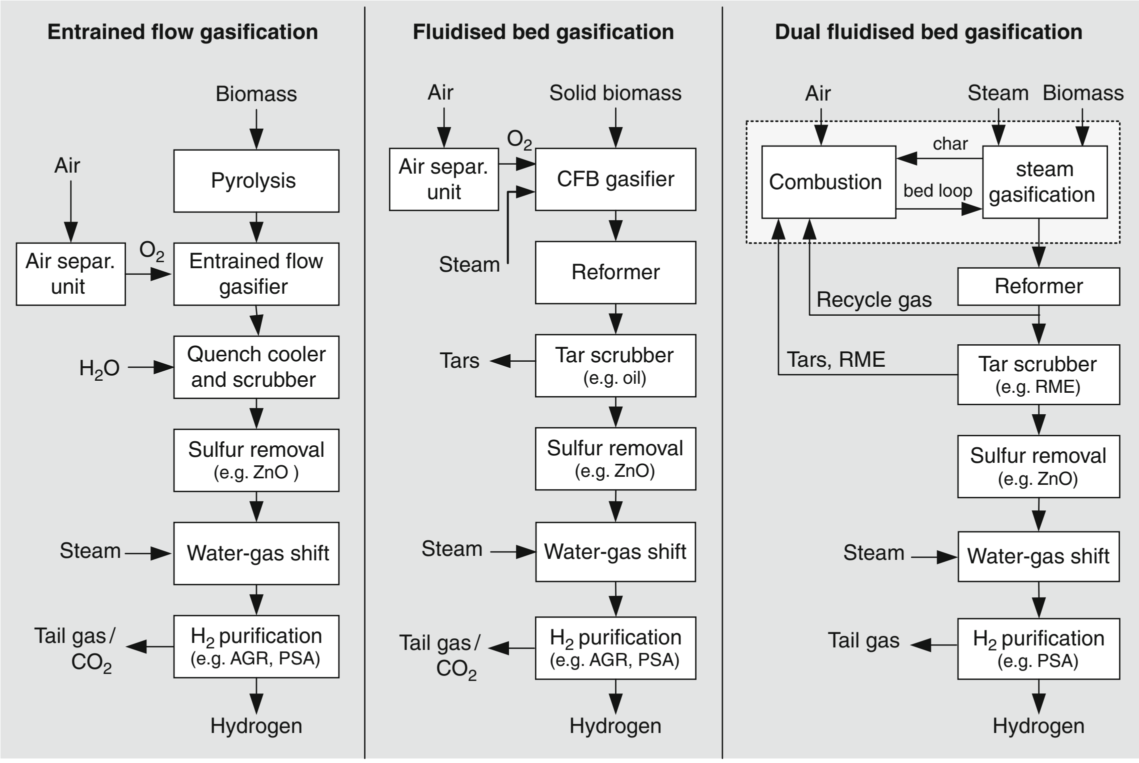
Recently, several large-scale biofuel production concepts based on entrained flow gasification were developed:

- The production of FT-Diesel from wood on a 45-MW scale with a staged gasification process has been demonstrated. The Carbo-V® process operates at 5 bar pressure. First, biomass is converted into tar-containing gas and charcoal by autothermal pyrolysis at 400–500°C. Second, the



Hydrogen from Biomass. Figure 7

Concept of biomass gasification plants for H_2 production



Hydrogen from Biomass. Figure 8
Biomass gasification process chains for H_2 production

pyrolysis gas is fed to an entrained flow gasifier operating at about 1,500°C. Third, charcoal is used for cooling the hot gas down to 900°C by a chemical quench process. The tar-free raw gas contains 30–40% of both H_2 and CO, balanced by CO_2 and H_2O 0.

- The bioliq® concept couples decentralized fast pyrolysis of straw with centralized entrained flow gasification of the resulting bioslurry. The fast pyrolysis is performed at ambient pressure and 500°C. The resulting char and pyrolysis oil are mixed to form slurry, which is transported to the gasification plant. There, it is fed to an entrained low gasifier operating at 30 bar and 1,200°C. The raw gas contains 30–40% CO and 20–30% H_2, balanced by N_2, CO_2, and H_2O. Favored product outputs are FT-Diesel and dimethyl ether, but H_2 production is considered as an option [20].

Fluidized bed gasification concepts, as shown in the middle part of Fig. 8, avoid the need for thermochemical biomass pretreatment. They have milder requirements regarding feedstock drying and sizing. Wood chips are a common fuel. Circulating fluidized bed (CFB) gasifiers are in some major aspects similar to entrained flow gasifiers. They use the same gasification agent – oxygen – and may operate pressurized.

An important difference compared to entrained flow gasifiers is related to ash. Because sintering and slag formation would create problems in terms of bed fluidization, the operating temperature is limited to roughly 1,000°C. At this operating condition, carbon conversion is around 95%. The limited carbon conversion is linked to a major problem of fluidized bed gasifiers: syngas quality suffers from tars. Since tars are not only impurities but also carry a significant amount of the chemical bound energy, tar treatment is crucial for the overall efficiency. A preliminary reforming of tars – and other hydrocarbons including methane, which is a major component in CFB raw gas – recovers the chemical energy to a large extent in form of CO and H_2 before a secondary tar removal process purifies the syngas. Further downstream processes are similar to those of entrained flow gasification based H_2 concepts. Differences in syngas composition and pressure level decide over the choice of the most suitable H_2 purification steps.

At present, CFB biomass gasification for H_2 is still under development. In Scandinavia, different demonstration projects deal with a slipstream H_2 train:

- In Värnamo (Sweden), the VVBGC demonstration project rebuilds a biomass-fueled IGCC plant into a synthesis gas producing unit with research and development [21].
- In Varkaus (Finland), a consortium is demonstrating a 12-MW syngas atmospheric oxygen/steam–blown CFB gasifier. The ultra clean gas (UCG) project focuses on syngas treatment to provide a reliable interface for different production targets, including H_2 [22, 23].

Dual fluidized bed gasification concepts, as shown in the right part of Fig. 8, cover process steps similar to those of CFB-based process chains. The major differences are: (1) the carbon conversion in the gasification chamber of roughly 75% leads to a smaller amount of syngas containing a higher percentage of H_2, (2) the duty of the water–gas shift is lower because the raw syngas composition provides a higher H_2/CO ratio due to the utilization of steam as gasification agent, and (3) tars are recycled to the combustion chamber because the segregation of gasification and combustion permits an easy handling of additional fuel.

Dual fluidized bed gasification concepts for the production of fuels and chemicals are mainly designed for methane output, often as a part of a polygeneration concept. But some concepts are as well optimized for H_2 production:

- A biomass gasification concept for the production of gaseous and liquid fuels has been introduced. A dual fluidized bed gasifier integrates a tar reformer utilizing the heat of the flue gases to boost the syngas H_2 content from 20% up to more than 45% [24, 25].
- The so-called Güssing model follows the strategy of decentralized, local energy production from all available renewable resources in a region. Here, an 8-MW_{th} FICFB gasifier (Fig. 6) is operated in CHP mode [26]. Additionally, the gasifier provides opportunities for testing different bed materials and for the use of a syngas side stream for process development purposes, for example, the development of methane and BtL fuel production downstream equipment.

Table 5 summarizes the characteristics of the presented biomass gasification process chains regarding H_2 production. Those process chains provide a basis for additional or pure H_2 output but rely on an initial design for other output options. H_2 production is specified as a potential output stream but has not been demonstrated on an industrial scale. From the view of process economics, a major drawback of current gasifier designs is the initial low H_2 content of the syngas. Recently, new processes with higher H_2 concentrations have gained interest. Particularly, a process called adsorption enhanced reforming (AER) provides good opportunities for commercial H_2 production from wood and herbaceous biomass [27] 0.

Case Study: The AER Process

The AER process is a new approach to improve dual fluidized bed gasification. Figure 9 describes the process principle. By using a bed material such as mineral calcite ($CaCO_3$) in a dual fluidized bed gasifier, a chemical loop is established where the bed material not only acts as a heat carrier but also adsorbs CO_2 and transports it from the gasification reactor to the combustion chamber where it is released.

CO_2 absorption in the gasifier part shifts the equilibrium of the water–gas shift reaction towards H_2 (Eq. 7). Hence, an H_2-rich product gas results with reduced CO_2 and CO concentration. The idealized sum reaction of the AER gasification – without considering the formation of methane and other secondary products – is given by Eq. 8 [28].

$$\begin{aligned} CH_xO_y + (2-y)H_2O + CaO &\rightarrow CaCO_3 \\ &+ (0.5x + 2 - y)H_2 \end{aligned} \tag{8}$$

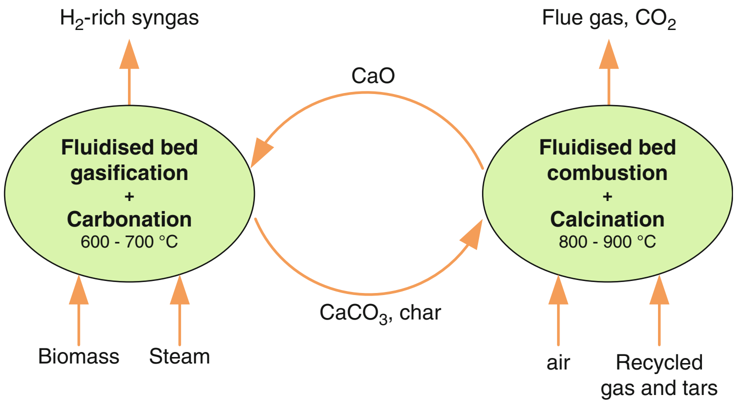
$$CaO + CO_2 \leftrightarrow CaCO_3 \quad \Delta H_R = -181.4\ \mathrm{kJ/mol} \tag{9}$$

The operating temperature of the gasification reactor is 600–700°C. According to Eq. 9, CO_2 adsorption (carbonation) is highly exothermal so that the heat

Hydrogen from Biomass. Table 5 Main characteristics of biomass gasification process chains for fuels and chemicals

	Dual fluidized bed	Circulating fluidized bed	Entrained flow
Gasification characteristics			
Reactor type	Dual fluidized bed	Fluidized bed	Entrained flow
Gasification agent	Steam	Oxygen (+steam)	Oxygen (+steam)
Heat supply	Bed loop	Autothermal	Autothermal
Temperature level (°C)	800–870	900–1,000	1,300–1,600
Pressure level (bar)	1	1–20	5–80
Residues	Ash	Ash	Slag
H_2 at outlet (vol%)	30–40	20–30	25–35
CH_4 at outlet (vol%)	9–12	10–15	<1
Required downstream processes for H_2 production			
Tar removal/conversion	Yes	Yes	No
Steam reforming	Yes	Yes	No
CO shift	Yes	Yes	Yes
H_2 purification	Yes	Yes	Yes
Plant characteristics			
Conversion efficiency[a]	50–60	50–60	45–55
Target scale (MW_{th})	10–100	30–200	250–1,000
Drawbacks	Tar load	Feeding system (pressurization), tar load	Biomass pretreatment, process control
Announced products	CH_4, BtL, polygeneration	BtL, CH_4, H_2, IGCC	BtL, CH_4, H_2

[a] Biomass to H_2 on LHV basis, estimation


H_2-rich syngas
Flue gas, CO_2
CaO
Fluidised bed gasification + Carbonation 600 - 700 °C
Fluidised bed combustion + Calcination 800 - 900 °C
$CaCO_3$, char
Biomass
Steam
air
Recycled gas and tars

Hydrogen from Biomass. Figure 9
Principle of AER gasification process

demand of the gasifier part is lower as compared to FICFB steam gasification. The fluidized bed material – calcite ($CaCO_3$) and char at this moment – is directed to the combustion chamber where char and additional fuel are burned to provide the calcination temperature well above 800°C. The heated bed material – now quicklime (CaO) – is recycled to the gasification chamber. Because the temperature difference of more than 200°C between gasifier and combustor is much higher than those 50°C of the FICFB process, it is necessary to reduce the bed material circulation rate. In order to do so, the particle size of the bed material is increased, and the primary fluidization in the combustor is decreased [29].

The desired properties on the bed material for AER gasification are:

- High mechanical stability to avoid dust problems in downstream equipment
- High CO_2 capacity and sufficient reaction rate of carbonation and calcination
- High catalytic activity towards tar conversion and CO shift reaction

Different limestones as a bed material have been tested with two biomass fuels, wood pellets, and straw pellets in a 100-kW_{th} process development unit [29]. The tested bed material consisted mainly of calcite coming from different mining companies. Clear variations in overall performance of wood gasification regarding gas composition and dust load have been observed:

- The H_2 concentration varies in a range from 59% to 70%.
- The tar content in raw syngas varies from 0.4 to 2.8 g/Nm^3 with most of the values lower than 1 g/Nm^3.
- The dust content is between 12 and 42 g/Nm^3.

In contrast to conventional dual fluidized bed steam gasification, the AER process provides favorable conditions for straw gasification. Straw is a difficult fuel for biomass gasification because of its low sintering point at 700–750°C, causing bed agglomeration in reactors operating at the temperatures above. The lower gasification temperature and the positive influence of CaO on the melting behavior of straw enable this widely available biomass resource for H_2 production. Compared to wood pellets, drawbacks are a higher dust load (~80 g/Nm^3) due to the high ash content, a higher tar content of 2–3 g/Nm^3, and a lower H_2 concentration of 52% [29].

The same 100-kW_{th} apparatus have been used to perform a comparison between FICFB and AER

Hydrogen from Biomass. Table 6 Comparison of wood gasification with conventional FICFB and AER process [30]

	FICFB	AER
Operating parameter		
Bed material	100 kg olivine	100 kg calcite
Steam/fuel ratio (kg/kg)	0.63	0.79
Gasification temperature (°C)	841	645
Combustion temperature (°C)	920	894
Syngas properties		
H_2 (vol%, dry)	39.1	73.9
CO (vol%, dry)	29.1	6.1
CO_2 (vol%, dry)	17.5	6.0
CH_4 (vol%, dry)	11.4	11.7
Hydrocarbons C2–C6 (vol%, dry)	2.9	2.3
Dust (g/Nm^3)	5.9	22.8
Char (g/Nm^3)	35.2	28.8
Tar (g/Nm^3)	3.5	1.4
H_2O (vol%)	33.2	51.4
Energy balance		
Fuel, wood pellets (kW)	123.5	123.5
Additional fuel, oil (kW)	26.6	9.6
Auxiliaries and steam (kW)	11.4	12.0
Syngas chemical power (kW)	105.6	66.5
Conversion efficiency[a] (%)	70.4	50.0
H_2 yield (l_{H2}/kg_{wood})	280	580

[a] Fuels to syngas, LHV

gasification regarding mass and energy balances [30]. The bed materials were olivine in case of FICFB gasification and calcite for the AER process. Oil was used as additional fuel. Table 6 provides an overview of the main results of those test runs. In contrast to conventional steam gasification, syngas properties of the AER process are more advantageous for H_2 production since the effort for downstream processing is much lower, but the conversion efficiency is comparatively low.

Future Directions

The market implementation of hydrogen from biomass is strongly influenced by the aimed utilization. The fields of use are significant:

- Chemical industry, replacing hydrogen from fossil fuel
- Transportation fuel
- Energy carrier for the storage and dissemination to decentralized applicants

The barrier for market access will be the lowest one in the first case because the market is established, the infrastructure is still available, and there is hardly any alternative to hydrogen as a raw material for processes realized within the chemical industry. The replacement of fossil fuel in the hydrogen production can generate an important environmental benefit. This can be done step by step if the production plants based on fossil fuels reached their end of the lifetime.

In the other market mentioned above (use as a transportation fuel and/or storage medium), a new infrastructure has to be installed before a hydrogen use in the large scale is possible. The creation of such a new infrastructure may be much more expensive than exclusive erection of the production plants. In addition, for these fields of application, renewable alternatives such as biomethane and liquid fuels still exist. It cannot be foreseen which position hydrogen will achieve within this competition field with the given fuel alternatives in years to come.

Some H_2 production methods from biomass are well developed, and several processes are demonstrated. Nevertheless, a commercial status could not be reached until now, neither based on the thermochemical nor on the biochemical or biological route:

- The process steps for the thermochemical hydrogen production based on biomass are well known and scientifically analyzed. Demonstration of the whole provision pathway in a near commercial size has to be the next implementation step. Basically, these processes have to show their economic performance.
- Biochemical pathways for the generation of hydrogen are still in the status of basic research. Therefore, these pathways might have the potential for an important contribution to bioenergy supply only if basic questions are solved based on increased research activities.

The viability of a future biomass to H_2 technology critically depends on the development of efficient, large-scale, cost-effective, and sustainable H_2 production systems. There are lots of activities ongoing, but so far, an economic breakthrough on the market cannot be seen. The years to come will show if this option has a realistic chance to contribute with a considerable share within the overall energy system [1, 31].

Bibliography

Primary Literature

1. Milne TA, Elam CC, Evans RJ (2001) H_2 from biomass – state of the art and research challenges. A report for the international energy agency agreement on the production and utilization of H_2; Task 16, H2 from carbon-containing materials. National Renewable Energy Laboratory, Golden. http://ieahia.org/pdfs/H2_biomass.pdf. Accessed July 2010
2. Hallenbeck PC, Benemann JR (2002) Biological hydrogen production; fundamentals and limiting processes. Int J Hydrogen Energy 27(11):1185–1193
3. Brunstermann R (2010) Entwicklung eines zweistufigen anaeroben Verfahrens zur kontinuierlichen Wasserstoff- und Methanerzeugung aus organischen Abfällen und Abwässer. Forum Siedlungswasserwirtschaft und Abfallwirtschaft, Universität Duisburg-Essen. Shaker, Aachen
4. Reith JH, Wijffels RH, Barten H (2003) Bio-methane and bio-hydrogen status and perspectives of biological methane and hydrogen production. Dutch Biological Hydrogen Foundation, The Hague. ISBN 9090171657
5. Petrovic T, Wagner HJ, Lente A, Widmann R (2005) Photobiologische Wasserstofferzeugung durch Mikroalgen – Beschreibung konkurierender Systeme zur H_2-Erzeugung. Stärkung der technologischen Position an den Ruhruniversitäten Teilprojekt 8 ef. Ruhr Förderkennzeichen 856569-T-170
6. Rechtenbach D (2009) Fermentative Erzeugung von Biowasserstoff aus biogenen Roh- und Reststoffen.

Dissertation am Institut für Abfallressourcenwirtschaft, Technische Universität Hamburg-Harburg. In: Stegmann R (ed) Hamburger Berichte 34. Verlag Abfall aktuell, Stuttgart

7. Krupp M (2007) Bio hydrogen production from organic waste and waste water by dark fermentation – a promising module for renewable energy production. Forum Siedlungswasserwirtschaft und Abfallwirtschaft, Universität Duisburg-Essen. Shaker, Aachen
8. Van Ginkel SW, Logan B (2005) Increased hydrogen production with reduced organic loading. Water Res 39:3819–3826
9. Hyvolution. http://www.fch-ju.eu/sites/default/files/documents/ga2010/pieternel_claassen.pdf
10. Modigell M, Schumacher M, Claassen PAM, Friedl A, Wukovits W (2006) Entwicklung eines zweistufigen Bioprozesses zur Produktion von Wasserstoff aus Biomasse. GVC IT, Wiesbaden
11. Zeidan AA, van Niel EWJ (2007) Developing a thermopilic hydrogen-producing co-culture for efficient utilization of mixed sugars. WHTC, Montecationi Terme
12. Levin DB, Pitt L, Love M (2004) Biohydrogen production: prospects and limitations to practical applications. Int J Hydrogen Energy 29(2):173–185
13. Hemmes K, de Groot A, den Uil H (2003) Bio-H2 Application potential of biomass related hydrogen production technologies to the Dutch energy infrastructure of 2020–2050. ECN-C-03-028. Energy Research Center, Petten
14. Schlegel HG (1997) Allgemeine Mikrobiologie, 1st edn. Thieme, Stuttgart. ISBN 3-13-444607-3
15. Hawkes FR, Dinsdale RM, Hawkes DL, Huss I (2002) Sustainable fermentative H_2 production: challenges for process optimization. Int J Hydrogen Energy 27:1339–1347
16. Benemann JR (2004) Biological production of hydrogen-methane mixtures for clean electricity production. In: 15th hydrogen energy conference, Japan
17. IHS (2005) CEN/TS 14961 Solid biofuels – fuel specification and classes. IHS, Englewood Cliffs
18. Hofbauer H, Veronik G, Fleck T, Rauch R (1997) The FICFB gasification process. In: Bridgwater AV, Boocock D (eds) Developments in thermochemical biomass conversion, vol 2. Blackie Academic & Professional, Glasgow, pp 1016–1025
19. Thrän D, Seidenberger T, Zeddies J, Offermann R (2010) Global biomass potentials – resources, drivers and scenarios. Energy Sustain Dev 14(3):200–205
20. Dinjus E, Kolb T, Dahmen N (2009) State of the art of the bioliq BTL process. In: European biomass conference and exhibition 2009, Hamburg
21. Albertazzi S, Basile F, Brandin J, Einvall J, Hulteberg C, Fornasari G et al (2005) The technical feasibility of biomass gasification for H_2 production. Catal Today 106:297–300. doi:10.1016/j.catted.2005.07.160
22. McKeough P, Kurkela E (2008) Process evaluations and design studies in the UCG project 2004–2007. VTT Research Notes 2434. http://www.vtt.fi/publications/index.jsp
23. Harju-Jeanty T, Nuortimo K, Hotta A, Coda-Zabetta E, Palonen J, Kokki S et al (2009) Innovative utilization of renewable energy sources to combat climate change. In: Fourth international bioenergy conference 2009, Jyväskylä
24. Paisley MA, Overend RP (2002) Verification of the performance of future energy resources' SilvaGas ® biomass gasifier – operating experience in the Vermont gasifier. Paper presented at the Pittsburgh coal conference, 24–26 September 2002, Pittsburgh
25. Paisley M (2007) Advanced biomass gasification for the production of biopower, fuels, and chemicals. AIChE, Salt Lake City
26. Hofbauer H, Rauch R, Loeffler G, Kaiser S, Fercher E, Tremmel H, et al (1998) Six years experience with the FICFB-gasification process. In: Twelfth European conference and technology exhibition on biomass for energy, industry and climate protection, Amsterdam
27. Koppatz S, Pfeifer C, Rauch R, Hofbauer H, Marquard-Moellenstedt T, Specht M et al (2009) H_2 rich product gas by steam gasification of biomass with in situ CO_2 absorption in a dual fluidized bed system of 8 MW fuel input. Fuel Process Technol 90(7–8):914–921. doi:10.1016/j.fuproc.2009.03.016
28. Marquard-Möllenstedt T, Sichler P, Specht M, Michel M, Berger R, Hein K et al (2004) New approach for biomass gasification to H_2. In: Proceedings of the second world conference and technology exhibition on biomass for energy, industry and climate protection, Rome, pp 10–14
29. Soukup G, Pfeifer C, Kreuzeder A, Hofbauer H (2009) In situ CO_2 capture in a dual fluidized bed biomass steam gasifier – bed material and fuel variation. Chem Eng Technol 32 (3):348–354. doi:10.1002/ceat.200800559
30. Pfeifer C, Puchner B, Hofbauer H (2009) Comparison of dual fluidized bed steam gasification of biomass with and without selective transport of CO_2. Chem Eng Sci 64(23):5073–5083. doi:10.1016/j.ces.2009.08.014
31. IEA Hydrogen Implementing Agreement (2006) Prospects for hydrogen from biomass, Annex 16 Subtask B, Final report

Books and Reviews

Bienert K (2009) Commercial scale BTL production on the verge of becoming reality. Biomass gasification seminar, Stockholm, 22–23 Oct 2009

Crotogino F, Hamelmann R (2007) Wasserstoff-speicherung in Salzkavernen. In: Tagungsband "14. symposium zur nutzung regenerativer energiequellen und wasserstofftechnik", Stralsund

Droste-Franke B, Bert H, Kötter A et al (2009) Brennstoffzellen und virtuelle kraftwerke. In: Gethmann CF (ed) Ethics of science and technology assessment, vol 36. Springer, Berlin, pp 43–131. http://www.springerlink.com/content/vu43p4/

DWV (2005) Deutscher Wasserstoff- und Brennstoffzellen-Verband e.V.; Press Release No. 5/05: Wasserstoff Spiegel 5/2005

DWV (2007) Deutschland schließt wieder zur Weltspitze auf–aber die Konkurrenz schläft nicht. Deutscher Wasserstoff- und Brennstoffzellen-Verband e.V.; Press Release No. 1/07 (22 Feb 2007)

Erdle EK (2001) Grundlagen, stand und perspektiven der brennstoffzellen-technik. In: VDI-Gesellschaft Energietechnik (ed) Stationäre brennstoffzellenanlagen: markteinführung Tagungsband. VDI, Düsseldorf, pp 3–14

FNR (ed) (2006) Wasserstoff aus biomasse, vol 25. Gülzower Fachgespräche, Gülzow

Gao M, Krishnamurthy R (2009) Hydrogen transmission in pipelines and storage in pressurized and cryogenic tanks. In: Gupta RB (ed) Hydrogen fuel – production, transport and storage. CRC-Press, Boca Raton, pp 341–379

Hoogwijk M, Faaji A, van Den Broek R, Berndes G, Gielen D, Turkenburg W et al (2003) Exploration of the ranges of the global potential of biomass for energy. Biomass Bioenerg 25:119–133. doi:10.1016/S0961-9534(02)00191-5

Jorde F (2006) Laboruntersuchungen zur fermentativen erzeugung von Biowasserstoff in den testsystemen sensomat und ATS unter Einsatz von Rein- und Mischkulturen. Diploma Thesis at the Institute of WasteRessourceManagement, TUHH, unpublished

Jurschik E-M (2007) Optimierung diskontinuierlicher versuche zur fermentativen bio-wasserstoffproduktion, sowie aufbau und inbetriebnahme eines bio-methanreaktors. Diploma Thesis at the Institute of WasteRessourceManagement TUHH, unpublished

Maddy J, Cherryman S, Hawkes FR, Hawkes DL, Dinsdale RM, Guwy AJ, Premier GC, Cole S (2003) Hydrogen 2003. Report Number 1, ERDF part-funded project entitled “A sustainable energy supply for Wales: towards the hydrogen economy.” University of Glamorgan, Pontypridd, Mid Glamorgan, Wales, ISBN 1-840540-90-7

Meyer M, Rechtenbach D, Stegmann R (2007) Biological production of H2 from organic raw and waste materials by fermentation with pure and mixed cultures. In: 15th European biomass conference, Berlin, 05 July–05 Nov 2007 (Proceedings CD)

Nandi R, Sengupta S (1998) Microbial production of hydrogen: an overview. Crit Rev Microbiol 24(1):61–84

Noike T, Takabatake H, Mizuno O, Ohba M (2002) Inhibition of hydrogen fermentation of organic wastes by lactic acid bacteria. Int J Hydrogen Energy 27:1367–1371

Pehnt M, Höpfner U (2009) Wasserstoff- und stoffspeicher in einem energiesystem mit hohen anteilen erneuerbarer energien: analyse der kurz- und mittelfristigen perspektiven. Kurzgutachten. IFEU -Institut für Energie und Umweltforschung, Heidelberg. http://www.bmu.de/files/pdfs/allgemein/application/pdf/ifeu_kurzstudie_elektromobilitaet_wasserstoff.pdf. Accessed 22 Feb 2010

Plath M (2006) Untersuchung und Optimierung der diskontinuierlichen erzeugung von biowasserstoff in dem testsystem ATS unter einsatz von klärschlamm und glukose. Project work at the Institute of WasteRessourceManagement, TUHH, unpublished

Rechtenbach D (2009) Fermentative erzeugung von biowasserstoff aus biogenen roh-und reststoffen. Hamburger Berichte 34, Abfallwirtschaft, T.U.Hamburg- Harburg, Germany, Verlag Abfall aktuell

Rechtenbach D, Meyer M, Stegmann R (2006) Fermentative production of bioH2 from organic raw and waste materials. In: Proceedings of the international conference ORBIT 2006 – biological waste management from local to global, 13–15 Sept 2006, Weimar; Kraft E, Bidlingmaier W, de Bertoldi M, Diaz LF, Barth J (Hrsg) Lombego systems & goldwiege. Visualle Projekte, Weimar, pp 869–877, ISBN 3-935974-09-4

Reiß T, Hüsing B (1993) Biologische wasserstoffgewinnung – forschungsperspektiven und technikfolgen. Verlag TÜV Rheinland, Köln. ISBN 3-8249-0150-1

Tanksale A, Beltramini JN, Lu GM (2010) A review of catalytic H_2 production processes from biomass. Renew Sustain Energy Rev 14(1):166–182. doi:10.1016/j.rser.2009.08.010

van Ginkel S, Sung S, Lay J-J (2001) BioH_2 as a function of pH and substrate concentration. Eviron Sci Technol 35:4726–4730

Zurawski D, Susanto AL, Stegmann R (2004a) Fermentative erzeugung von bio-wasserstoff aus biogenen roh- und reststoffen. In: Proceedings (CD-ROM) des 10. Internationalen kongresses für nachwachsende rohstoffe und pflanzenbiotechnologie NAROSSA, Magdeburg, 07–08 June 2004

Zurawski D, Susanto AL, Stegmann R (2004b) Sind bioabfälle und energiepflanzen zur biologischen wasserstofferzeugung geeignet? In: Fricke K, Kosak G, Wallmann R, Fischer J, Vogtmann H (eds) Schriftenreihe des ANS, 65. Informationsgesprächs des ANS e.V. “EEG und Emissionshandel – Neue Chancen für Biomassenutzung und Abfallwirtschaft”, 06–07 Dec 2004, Braunschweig, pp 189–200. ISBN 3-935974-06-X

Zurawski D, Meyer M, Stegmann R (2005) Fermentative production of bioH_2 from biowaste using digested sewage sludge as inoculum. In: Cossu R, Stegmann R (eds) Proceedings of Sardinia 2005 – tenth international waste management and landfill symposium. CISA Environmental Sanitary Engineering Centre, Santa Margerita di Pula, Cagliari, 03–07 Oct 2005

Hydrogen Production from High-Temperature Fuel Cells

JACOB BROUWER, PERE MARGALEF
National Fuel Cell Research Center, University of California, Irvine, CA, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Background
Key Principles of Poly-Generation of H_2 with HTFC

Cycle Configurations for Poly-Generating HTFC
Comparative Efficiency Analyses
Hydrogen Transport Impacts on Comparative Analyses
Discussion of Scientific Basis for Observed Performance
World's First Poly-Generating HTFC System
Future Directions
Bibliography

Glossary

Distributed energy resources (DER) Small-scale power generation, energy conversion, storage, and/or control technologies (typically in the range of 3–10,000 kW) that are installed and operated close to the energy demand (e.g., a home or business). These resources can provide an alternative to or an enhancement of the existing electric power and thermal energy systems. Among the main advantages of the DER are the energy saving associated with the transport of energy from the centralized plants to the points of use and the potential to poly-generate power, heating, cooling, and/or fuels close to the point of use.

Distributed generation (DG) According to the California Energy Commission [7], distributed generation (DG) comprises small-scale power generation technologies (typically in the range of 3–10,000 kW) located close to where electricity is used (e.g., a home or business) to provide an alternative to or an enhancement of the existing electric power system. Among the main advantages of the distributed generation are the energy saving associated with the transport of energy from the centralized plants to the points of use and the potential for cogeneration of heat and power.

High-temperature fuel cells (HTFC) HTFC are electrochemical conversion devices that produce electricity directly from the chemical potential difference between a fuel and oxidant. There are two common types of high-temperature fuel cells: (1) molten carbonate fuel cells (MCFC) and (2) solid oxide fuel cells (SOFC).

The *molten carbonate fuel cell* uses a carbonate electrolyte, which is generally a mixture of lithium and potassium carbonates (salts). At the high operating temperature (typically 550–650°C) the alkali carbonates become a highly conductive molten salt, with $CO_3{}^{2-}$ ions providing ionic conduction. Anode materials are typically Ni–Cr/Ni–Al alloys and cathode materials are comprised of lithiated NiO [2]. The fuel cell operation is the result of a complex conjunction of physical, chemical, and electrochemical processes that together oxidize fuel and reduce oxidant in separate compartments to produce electricity, heat, and redox reaction products. The anode and cathode half reactions and overall electrochemical reaction are [3]

$$\text{Cathode}: 0.5O_2 + CO_2 + 2e^- \rightarrow CO_3^{2-} \quad (1)$$

$$\text{Anode}: H_2 + CO_3^{2-} \rightarrow H_2O + CO_2 + 2e^- \quad (2)$$

$$\begin{aligned}\text{Overall}: H_2 + 0.5O_2 + CO_{2,\text{cat}} &\rightarrow H_2O \\ &+ CO_{2,\text{an}} + \text{electricity} + \text{heat}\end{aligned} \quad (3)$$

Note that carbon dioxide must be supplied to the cathode as well as oxygen. The CO_2 is converted to carbonate ions which provide the means of ion transfer between the cathode and the anode. Therefore, there is a net transfer of CO_2 from cathode to anode. There are different ways of CO_2 recycling. The most common method feeds the anode exhaust gas to an anodic gas oxidizer (AGO), which converts any unused hydrogen or carbon monoxide into water and CO_2. A portion of the exhaust gas from the anodic gas oxidizer is then mixed with fresh air and fed into the cathode inlet. This process also serves to preheat the reactant air, burn the unused fuel, and bring the waste heat into one stream for use in a bottoming cycle [4].

The *solid oxide fuel cell* uses an oxide ion-conducting ceramic material as the electrolyte. Since only two phases (solid and gas) are required and CO_2 recirculation is not required, these systems are conceptually simpler than other fuel cells [5]. As in the molten carbonate case, no precious metal catalysts are needed due to the high operating temperatures (700°C–1,000°C). SOFC have typically used materials sets based upon a yttria-stabilized zirconia (YSZ) electrolyte comprised of 8–10% Y_2O_3 in ZrO_2 [2]. Above 700°C, YSZ becomes a conductor of oxygen ions (O^{2-}). The negatively charged ion (O^{2-}) is transferred from the cathode to the anode [6]. The half-reactions produce water in the anode as follows:

$$\text{Anode}: H_2 + O^{2-} \rightarrow H_2O + 2e^- \quad (4)$$

$$\text{Cathode}: 0.5O_2 + 2e^- \rightarrow O^{2-} \quad (5)$$

$$\text{Overall}: H_2 + 0.5O_2 \rightarrow H_2O + \text{electricity} + \text{heat} \quad (6)$$

The remaining SOFC fuel from the anode reactions is typically oxidized by the remaining oxygen from the cathode in a combustor to produce heat. This heat is used in the fuel cell plant to preheat the inlet streams and/or overcome the endothermicity of steam reforming reactions elsewhere in the system.

Hydrogen separation When hydrogen is produced by thermochemical processes such as steam methane reformation (SMR), the output gas is a hydrogen-rich gas (i.e., reformate) that contains water vapor, carbon monoxide, carbon dioxide, and other trace gases. Therefore, hydrogen must be separated from the reformate gas and purified to reach the specific requirements for hydrogen use.

Hydrogen separation in refineries has been traditionally done by established technologies such as pressure swing adsorption (PSA), selective permeation processes using polymer membranes, or cryogenic separation processes. Each process is based upon different separation principles. Economic aspects and other project considerations such as process flexibility, reliability, and scalability have to be taken into account to decide the hydrogen separation method [10].

Besides the traditional methods, alternative techniques for hydrogen separation are being developed. These include the Electrochemical Hydrogen Separation method (EHS) and Hydrogen Separation Membrane Reactor (HSMR) [11].

Parasitic loads In power generation devices, parasitic loads are those loads associated with the normal operation of the system that are required to sustain the normal operation of the system in a safe and reliable manner. Typical parasitic loads are associated with the preparation of the reactants, handling of products, exchanging of heat and flows, and thermal management. The power required for these parasitic loads must be produced by the same generator and subtracted from the generator power to produce the net system power output. Ideally, parasitic loads should be minimized in order to achieve higher system efficiencies.

Poly-generation Poly-generating systems include energy conversion systems that convert fuel chemical energy into multiple useful forms of energy or power. For instance, a specific poly-generation system instance that produces only two products is a combined heat and power (CHP) system that generates electrical and thermal power from fuel in a single, integrated system. Poly-generating systems are typically comprised of a number of individual components, including for example a heat engine, generator, heat recovery equipment, and electrical interconnection hardware that are configured into an integrated whole system. The primary energy conversion device (i.e., the prime mover) is typically used to identify the type of poly-generating system. Prime movers include reciprocating engines, gas or steam turbines, and fuel cells that can produce electrical and thermal power from a variety of fuels, including natural gas, coal, and biofuels. Thermal energy from the system can be used in direct process applications or indirectly to produce steam, hot water, hot air for drying, or chilled water for process cooling [8]. High-temperature fuel cells, in particular, can be designed and operated to poly-generate electricity, heat, and useful chemicals (such as hydrogen) in a variety of configurations [9].

Steam methane reformation (SMR) Steam methane reforming is a mature industrial technology that is typically used for hydrogen production. The basic reforming reaction for methane is

$$CH_4 + H_2O \rightarrow 3H_2 + CO \quad (7)$$

SMR is an endothermic reaction so that heat must be provided to drive the reaction forward to produce hydrogen. The overall process for hydrogen production typically also includes the water–gas-shift reaction defined below.

Synergy Synergy describes the complementary interaction of processes to create an outcome that is in some way of more value than the sum of the individual values that would otherwise have been produced by the individual processes operating independently.

Water–gas shift (WGS) The reaction of carbon monoxide with water to produce hydrogen and carbon dioxide as follows:

$$CO + H_2O \rightarrow CO_2 + H_2 \quad (8)$$

The WGS reaction is reversible and exothermic.

Well-to-tank (WTT) efficiency Well-to-tank efficiency is defined as the total energy of a ready-to-be-used fuel (i.e., in the tank of the vehicle), divided by the total energy required for the extraction, preparation, transport, and dispensing of the same fuel. This efficiency calculation is widely used in the automotive sector to compare the energy intensity of different fuels. On the other hand, well-to-wheel (WTW) efficiency accounts for the inefficiencies of the vehicle propulsion system. WTT are higher than WTW efficiencies.

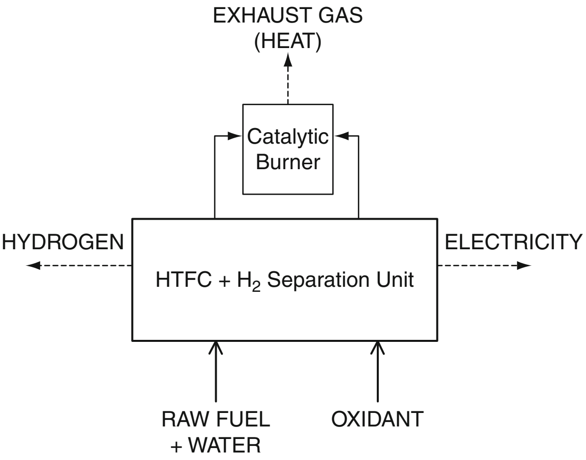
Definition of the Subject

Hydrogen is a likely energy carrier of the future due to the absence of carbon, low emissions when converted in various end-use technologies, and ability to be cleanly and efficiently produced from various domestic primary energy sources. In 2003 the Federal government launched the Hydrogen Fuel Initiative with a total budget of $1.2 billion over 5 years in order to accelerate research and development of fuel cell technologies [12]. Importantly, major automobile manufacturers are operating fuel cell vehicles that run on pure hydrogen gas, and several fuel cell buses are in operation in major cities around the world. Companies such as Shell, Air Products and Chemicals, Chevron, and Air Liquide are developing hydrogen production, distribution, and dispensing technologies for hydrogen vehicles along with strategies to deploy them. Moreover, the state of California intends to reduce the carbon content of transportation fuels through the Low Carbon Fuel Standard. Additional legislation in the areas of energy and climate, such as that contained in California laws entitled AB32, AB1493, SB76, and SB1368, highlight California's commitment to ensuring low greenhouse gas emissions for both electricity and transportation fuels [12]. Many other places around the world, including Germany and fellow European Union countries, Japan, China, and Korea, are advancing hydrogen fueling and vehicle technology. Therefore, developing a highly efficient, low-emission, and economically viable hydrogen production and delivery methodology is of great importance from environmental, social, economic, and political perspectives.

The production of hydrogen from high-temperature fuel cells is accomplished by a synergistic integration of fuel processing, electrochemical conversion in a fuel cell, and hydrogen separation that leads to an ability to locally produce and deliver hydrogen with ultralow criteria pollutant and greenhouse gas emissions. The resulting system comprised of a high-temperature fuel cell that produces electricity, heat, and hydrogen fuel uses a hydrogen separation unit (HSU) that separates and purifies the hydrogen from the anode off-gas, and is typically referred to as a poly-generating or tri-generating fuel cell. Figure 1 shows a basic schematic of a poly-generating HTFC.

Introduction

Since it is not naturally occurring, hydrogen must be produced from water via electrolysis, photolysis, or thermal splitting, or from hydrocarbon fuels (e.g., natural gas, coal, biogas, biomass) via reformation or gasification. Currently, hydrogen is typically produced at large centralized steam methane reformation (SMR) plants from natural gas. SMR is a convenient and



Hydrogen Production from High-Temperature Fuel Cells. Figure 1
Basic schematic of a poly-generating HTFC

cost-effective method that has been implemented to produce a substantial commercial supply for petroleum refining, metals processing, and other industrial uses. Not only hydrogen production (e.g., SMR plants and a host of other cost-effective and environmentally sensitive production technologies) but also energy-efficient and environmentally sensitive technologies for transport, distribution, and dispensing of hydrogen will be needed in the future [13].

Distributed generation of hydrogen, while typically less efficient than centralized generation, is foreseen by many as a viable strategy to produce hydrogen due to the savings associated with transport and distribution of the hydrogen, which can be energy and emissions intensive. Significant efforts to develop small-scale hydrogen production plants have led to commercialization of relatively affordable systems. Ogden [14] reviewed the development and commercialization status of various types of small-scale reformers.

Conventional steam methane reformers consisting of long catalyst-filled tubes that operate at high pressures (15–25 atm) and high temperatures (850°C) have been successfully scaled down to units that produce as little as 10–100 kg/h. However, at such small sizes, relative capital costs are too high to compete with large-scale hydrogen production. In addition, the footprint of these systems may be too large to be placed at conventional fueling stations. For this reason, more compact and inexpensive designs have been, and are being, developed by many manufacturers. The average hydrogen production efficiency reported for good systems ranges between 60% and 77% (on a LHV basis) [14]. The main technical challenges of distributed reformers include relatively larger system heat losses at small scale and inability to produce hydrogen on demand that may require large hydrogen storage capabilities, especially during the early stages of hydrogen vehicle deployment.

Background

As indicated above, most hydrogen is produced today from fossil fuels in large SMR plants and is used at or near the production site [13]. As fuel cells and hydrogen vehicles become more widely used, an entire infrastructure for the distribution and dispensing of hydrogen will be needed with the additional goal of producing and delivering hydrogen to consumers in an environmentally sensitive manner [15].

An integrated energy future that meets transportation fuel and stationary power and thermal energy demands could be made from the same primary energy sources. For example, the primary feedstock currently used for the production of hydrogen (i.e., natural gas) is also a significant primary energy source for electricity production. As both energy sectors (transportation and electricity generation) come to rely on the same primary energy sources, there are significant opportunities to integrate them, which can introduce profound changes in how our energy is converted and distributed. The integration of these two energy sectors at the level of a production plant via poly-generation of hydrogen, heat, and electricity could lead to lower prices for both transportation and stationary applications and ultimately enhance overall efficiency and flexible use of diverse resources [9].

Electricity and hydrogen poly-generation can be accomplished at large or at small scale. For the large-scale approach, hydrocarbon feedstocks such as coal, natural gas, or biomass can be converted via thermochemical processes to a syngas. Syngas is comprised of hydrogen, water, carbon monoxide, and carbon dioxide and can be used to generate electricity in a fuel cell, steam cycle, gas turbine, or combination of these technologies (i.e., combined cycle). Since the syngas produced has a high concentration of hydrogen, part of the gas stream can be diverted and the hydrogen can be separated from the rest of the gas to produce high-purity hydrogen for fuel cell vehicle use.

For the small-scale approach, poly-generating fuel cells represent a viable technology to produce hydrogen, electricity, and heat on demand in a distributed fashion. Excess heat released during the electrochemical fuel cell reactions can be used to produce hydrogen through steam reformation of hydrocarbon fuels. The highly synergistic nature of the poly-generating concept at high temperatures leads to higher production efficiencies compared to conventional hydrogen and electricity generation [16]. If successfully developed, poly-generating HTFC that produce electricity, heat, and hydrogen from a variety of hydrocarbon fuels will provide high efficiency and low emissions distributed hydrogen production and delivery. In addition, such a concept could aid fuel cell market viability, stakeholder confidence, and energy security and sustainability together with emissions reduction [17].

Key Principles of Poly-Generation of H_2 with HTFC

The key principles that apply to high-temperature fuel cell systems that poly-generate electricity, heat, and hydrogen are as follows:

1. Fuel flexibility of high-temperature fuel cells
2. High-temperature fuel cell–fuel processing relationship
3. Low entropy associated with exchanging heat at similar temperature
4. Fuel utilization concept requires excess fuel processing
5. Lower fuel utilizations lead to higher electrochemical efficiencies
6. Endothermicity of hydrocarbon reformation provides needed cell cooling
7. Production of excess hydrogen produces higher fuel cell efficiency
8. Chemical synergy associated with reactant/product interactions
9. All synergies lead to lower fuel cell system parasitic losses
10. Electrochemical and catalytic processes are inherently low emissions
11. Processes are efficient even at small size enabling distributed production
12. Hydrogen transport involves emissions and energy penalties

A brief description of each of these key principles is presented in this section.

Fuel Flexibility of High-Temperature Fuel Cells

High-temperature fuel cells, such as the molten carbonate fuel cell (MCFC) and solid oxide fuel cell (SOFC), can be operated on a variety of hydrocarbon fuels including natural gas, digester gas, landfill gas, coal and biomass synthesis gases. MCFC and SOFC technologies comprise specific characteristics that make them especially amenable to operation on such hydrocarbon fuels. These characteristics include the use of an oxidizing ion in the electrochemical reactions (${CO_3}^{2-}$ for MCFC and O^{2-} for SOFC) and high-temperature operation (550–650°C for MCFC and 700–1,000°C for SOFC) which promotes sufficiently rapid chemical and electrochemical reactions.

High-Temperature Fuel Cell: Fuel Processing Relationship

Solid oxide fuel cells (SOFC) and molten carbonate fuel cells (MCFC) generate electricity and heat through exothermic electrochemical reactions. Oxidation of hydrogen takes place in the anode compartment, which overall thermodynamics are described by the global reaction:

$$H_2 + 0.5O_2 \rightarrow H_2O \quad \Delta\bar{h}_f = -241.83\,\mathrm{KJ\,mol^{-1}} \tag{9}$$

The electrochemical oxidation reactions that convert hydrogen to water in the SOFC and MCFC anode compartments are, respectively:

$$H_2 + O^{2-} \rightarrow H_2O + 2e^- \tag{10}$$

$$H_2 + CO_3^{2-} \rightarrow H_2O + CO_2 + 2e^- \tag{11}$$

These reactions are complemented by corresponding oxidant reduction reactions in the cathode compartment, ion transport, and electron flow to perpetuate the electrochemical reactions. Since electricity is produced by these electrochemical reactions at less than 100% efficiency, the remaining portion of the enthalpy of reaction produces heat as reactants are converted to products.

HTFC system designs typically incorporate fuel processing with electrochemical conversion. The global fuel processing reaction includes steam methane reforming and water–gas-shift reactions and is described by:

$$CH_4 + 2H_2O \rightarrow 4H_2 + CO_2 \quad \Delta\bar{h}_f = -165\,\mathrm{KJ\,mol^{-1}} \tag{12}$$

The heat generated by the electrochemical reactions of Eq. 9 tends to be greater than the heat required by the endothermic fuel processing reactions of Eq. 12 for the amount of electricity produced [6]. Surplus heat is typically used to preheat the fuel and oxidant streams before they enter the fuel cell and to produce the steam required for system operations. Therefore, more hydrocarbon fuel than that required for the electric power generation could be processed in a HTFC, creating a hydrogen-rich stream that could be subsequently purified and delivered to the point of use without the need of an additional reformer.

Low Entropy Associated with Exchanging Heat at Similar Temperature

To proceed at a sufficiently fast rate and to completion (i.e., consume all the hydrocarbon fuel to produce hydrogen-rich syngas), the fuel processing reactions (e.g., steam reformation reaction of Eq. 12) must occur at a reasonably high temperature. Irreversible heat transfer processes between the exothermic fuel cell reactions (i.e., heat source) and the endothermic reforming reactions (i.e., heat sink) increase the entropy generation of the system. This irreversible heat transfer must be provided to a typical reformer reactor in a fuel cell system, using either fuel combustion or fuel cell exhaust heat at temperatures above those at which the reformation reactions proceed. However, the operating temperatures of HTFC (550–650°C for MCFC and 700–1,000°C for SOFC) are similar to the temperatures at which hydrogen production rate and reaction completion are acceptable for steam methane reformation processes (i.e., ~700°C). Therefore, heat can be transferred from the exothermic to the endothermic process with minimal (or zero) temperature difference leading to low entropy generation for heat exchange, resulting in overall high efficiency.

Fuel Utilization Concept Requires Excess Fuel Processing

Fuel and oxidant utilization factors refer to the fractions of the total fuel and oxidant flowing through the anode and cathode compartments that are consumed to generate electricity. If fuel and/or oxidant species concentrations become too low, the chemical potential difference becomes unable to sustain a voltage difference. The Nernst potential (V_{Nernst}) describes the relationship between voltage and reactant and product concentrations that applies to a fuel cell

$$V_{\mathrm{Nernst}} = E_0 + \frac{RT}{2F} \ln \frac{x_{H_2} x_{\mathrm{O}_2}^{\frac{1}{2}}}{x_{H_2\mathrm{O}}} \tag{13}$$

where E_0 is the ideal reversible potential, F is Faraday's constant [96,487 kC/kmol], R is the gas constant, T is the operating temperature, and x_k is the concentration of the species k [6].

Since fuel cell electrodes are good electronic conductors (they act as equipotential surfaces), low reactant concentrations anywhere in the anode or cathode compartments leads to voltage that cannot be sustained. Therefore, fuel utilization ($\mathrm{U_f}$) and oxygen utilization ($\mathrm{U_{O_2}}$) must be always lower than 100%. In other words, the amount of fuel and oxidant fed into the fuel cell must always and continuously be greater than the amount consumed within the stack to produce electricity even when hydrogen is not produced.

Lower Fuel Utilizations Lead to Higher Electrochemical Efficiencies

In operational fuel cell systems, cell voltage increases as fuel utilization decreases, which results in higher electrochemical efficiencies. The electrochemical efficiency is defined as the actual cell voltage (V_{cell}) divided by the thermodynamically possible cell voltage ($V_{\max}$), or

$$\eta_{\mathrm{cell}} = \frac{V_{\mathrm{cell}}}{V_{\max}} \tag{14}$$

where $V_{\max}$ is defined as the maximum theoretical voltage that would be achieved if all the energy from the hydrogen fuel were transformed into electrical energy [16], or

$$V_{\max} = \frac{-\Delta \bar{g}_f}{n_i F} \tag{15}$$

where $\Delta \bar{g}_f$ is the Gibbs free energy, n_i is the number of electrons transferred during the electrochemical oxidization of fuel species i ($\mathrm{n_i} = 2$ for H_2 and CO, $\mathrm{n_i} = 8$ for CH_4), and F is Faraday's constant.

From the Nernst voltage equation Eq. 13, it can be observed that higher species concentrations lead to higher cell voltages. At lower fuel utilizations, reactant species concentrations are higher along the electrode channels. Thus, it can be stated that lower fuel utilization results in higher cell voltages that raise the electrochemical efficiency of the fuel cell as described by Eq. 14.

An alternative concept that provided insight into the effects of the fuel utilization on the electrochemical efficiency relates to the mixing of products with reactants to dilute the electrochemically active species concentrations. Water product mixes with the anode gas as it is formed by the electrochemical reactions in an MCFC or SOFC. The concentration of water product in the anode increases as the fuel utilization factor increases. Mixing processes are irreversible so they produce entropy. As a result, a completely reversible fuel cell operation would be only approached as $\mathbf{U_f} \rightarrow 0$. In other words, irreversibilities increase as the utilization factor increases [18].

Consequently, one desires to operate a fuel cell at the lowest possible utilization factor to maximize the cell voltage and reduce mixing irreversibilities. However, unless one recycles or otherwise uses the anode off-gas then one alternatively desires high utilization to achieve high overall stack electrical efficiency. Typically, fuel utilization is selected in the 70–90% range to balance these considerations.

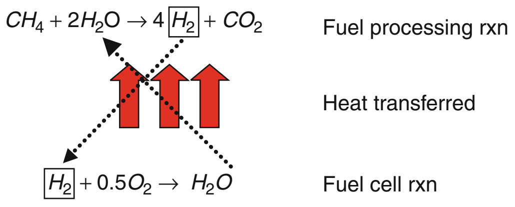
One of the key principles of poly-generating fuel cells is related to the capability of operating the fuel cell at lower fuel utilizations without compromising the overall system electrical efficiency. As already stated, hydrogen that is not electrochemically oxidized will be separated and considered as a valuable system output. Therefore, inasmuch as additional hydrogen is separated and used, there is the opportunity to take advantage of the higher electrochemical efficiency that is a natural by-product of the lower fuel utilizations required for hydrogen production.

Endothermicity of Hydrocarbon Reformation Provides Needed Cell Cooling

The fact that fuel cells must be cooled to maintain steady state operating temperature allows for synergistic coupling of endothermic steam reforming with exothermic fuel cell electrochemical reactions to yield unprecedented efficiency. Adding more endothermic reforming reactions for the same amount of electrochemical reactions results in a reduced requirement for air cooling of the cell. The synergistic performance benefits related to these lower cooling requirements are caused by: (1) the additional endothermic reformation at lower fuel utilization that provides additional cooling to the fuel cell stack; and (2) less electrochemical heat generated per mol of input fuel due to the higher voltages achieved at lower fuel utilizations. Both phenomena reduce the auxiliary power associated with forcing air through the cathode compartment, which is the primary means of otherwise removing heat from the fuel cell.

Chemical Synergy Associated with Reactant/Product Interactions

High-temperature fuel cells operate at high enough temperatures to convert methane into fuels that are more amenable to electrochemical oxidation such as hydrogen and carbon monoxide within the anode compartment. This concept is known as internal reformation (IR) and represents one of the key features of high-temperature fuel cells that can yield exceptionally high electrical efficiencies when compared with low-temperature fuel cells that process fuel in a separate reactor (i.e., external reformation). The main synergy associated with IR is related to the direct consumption of the hydrogen product of the reforming reaction by the electrochemical fuel cell reactions that produce water, a reformation reactant, which all occurs at the same time and physical location [17]. The immediate consumption of products and provision of reactants by the fuel cell reactions drives the reformation reaction forward and assures reaction completeness. Continual production of hydrogen and consumption of water by reformation also directly facilitates electrochemical reaction progress. This chemical synergy is introduced in addition to the heat exchange synergy between exothermic fuel cell reactions and endothermic fuel processing reactions with minimum losses since both reactions occur at the same time and place. Figure 2 shows the main


$CH_4 + 2H_2O \rightarrow 4\boxed{H_2} + CO_2$
Fuel processing rxn
Heat transferred
$\boxed{H_2} + 0.5O_2 \rightarrow H_2O$
Fuel cell rxn

Hydrogen Production from High-Temperature Fuel Cells. Figure 2
Internal reformation chemical and thermal synergy mechanisms

reactions, chemical exchange synergy, and heat transfer processes occurring during internal reformation in high-temperature fuel cells.

All Synergies Together Work in the Same Direction to Increase Fuel Cell System Efficiency

At low fuel utilizations, the ratio between endothermic heat from the fuel processing reactions and exothermic heat from the electrochemical reactions increases, leading to a reduction of the auxiliary power required to circulate cooling air through the cathode. In addition, thermodynamic analyses demonstrate that cooling required per mol of fuel input drops at lower fuel utilizations due to the higher operational voltages at low fuel utilizations [16]. The design and operating changes that produce these synergies require changes in the same direction. Therefore, all of the described synergies of poly-generating high-temperature fuel cells can work together to result in higher fuel cell efficiencies compared to conventional fuel cell system designs that do not poly-generate. There remain design challenges, such as managing the fuel cell temperature gradient in poly-generating systems, but, the overall design and operating changes required for poly-generation work together to improve efficiency.

Integration with Hydrogen Separation Produces Higher System Efficiency

Fuel cells are by nature required to produce somewhat more hydrogen than is electrochemically converted in the anode compartment. However, operating at lower fuel utilization leads to higher electrochemical efficiency and the chemical synergies of internal reforming, and the endothermicity of fuel processing reactions leads to lower parasitic losses due to reduced air blower power. These features by themselves would not be beneficial to overall system efficiency until and unless they are integrated into a system design that produces a significant amount of product hydrogen leading to remarkably higher overall efficiency. This is accomplished by using the above principles in combination with hydrogen separation technology that is integrated with the system design requirements in a manner that exchanges more heat for useful electrochemical and hydrogen production purposes.

Electrochemical and Catalytic Processes Are Inherently Low Emissions

The inherent electrochemical and catalytic nature of high-temperature fuel cells and poly-generating fuel cells yields ultralow criteria pollution emissions.

- Thermal NO_x requires high-temperature combustion in air whereas in HTFC, all of the fuel is processed and converted either electrochemically or by catalytic processes at low temperature.
- SO_x cannot be produced since sulfur is removed from the fuel before entering the system.
- Particulate matter (PM) is not produced due to high water content, catalyst presence, and low-temperature conditions extant in the anode compartment, followed by catalytic and/or low-temperature oxidation of the anode off-gas.
- CO is amenable to the electrochemical oxidation to release electrons in the anode compartment and the remainder is consumed by catalytic and/or low-temperature oxidation of the anode off-gas.
- Hydrocarbons are usually converted into methane in a pre-converter reactor before entering the fuel cell stack and the remainder is consumed by catalytic and/or low-temperature oxidation of the anode off-gas.
- Air toxics, such as higher hydrocarbons, aldehydes, and alcohols, are not present or are removed from the incoming fuel and typically no air toxic compounds are produced by any of the processes involved in a fuel cell system. If any air toxics would be produced then they would likely be destroyed at the high-temperature catalytic conditions of the anode compartment or anode off-gas oxidizer.

Processes Are Efficient Even at Small Size Enabling Distributed Production

Heat engine energy conversion is driven by a temperature difference while fuel cell energy conversion is driven by a chemical potential difference. The efficiency of any heat engine is limited by the efficiency that could be achieved by the reversible cycle operating upon a temperature difference, known as the Carnot efficiency [19]. On the other hand, fuel cells are limited by the Nernst equation and chemical potential difference that can be established in the cell, since they convert the chemical potential difference directly into electrical energy [6]. Heat engines for electricity production also have additional losses associated with the conversion of mechanical energy (e.g., piston movement, spinning turbine) to electricity through cranks, gears, and a generator.

Heat engines have reduced efficiency at reduced size because surface-to-volume ratios increase leading to higher percentage heat losses and inability to sustain the high cycle temperature for smaller engines. This leads to an inherent inability to establish a high temperature difference leading to a lower efficiency limit (i.e., lower Carnot efficiency). In addition, small-scale heat engines have proportionally larger friction, non-isentropic compressions and expansion, and other losses in comparison to larger heat engines. As a result, we currently tend to produce most of our power in very large central power plants based upon the heat engine – smaller plants are inherently less efficient.

On the other hand, fuel cell power production and efficiency depends upon the establishment of a chemical potential difference which can be established in equal difference regardless of the size of the cell.

Hydrogen Transport Involves Emissions and Energy Penalties

Today, most hydrogen is produced from natural gas in large steam methane reformation (SMR) plants in a centralized fashion. To produce hydrogen at large scale in centralized reformation plants is more efficient than producing hydrogen at small-scale reformation plants.

The main steps of a well-to-tank hydrogen supply chain may be defined as production, treatment, distribution, storage, and dispensing. Each step can be accomplished with a variety of different technologies which may be more or less energy intensive and environmentally friendly [20]. For long distances, transportation of liquefied hydrogen by diesel truck is the most common strategy. However, liquefaction of hydrogen is the most energy intensive process and it makes sense only for very long delivery distances. Shorter distance hydrogen transport may be resolved by the transportation of compressed hydrogen which takes less energy than liquefaction, but delivers less hydrogen per unit truck volume. Hydrogen can also be transported via high-pressure pipeline, which is one of the least energy intensive methods. But, in comparison to other gaseous or liquid fuels, hydrogen pipeline delivery is more energy intensive [26]. In all of these cases, the relatively low volumetric energy density of hydrogen leads to relatively large energy and emissions penalties associated with hydrogen transport and delivery. These penalties can result in mediocre well-to-tank efficiencies compared to other fuels. These penalties can be averted by the production of hydrogen in a distributed fashion with a technology that scales down with high efficiency, such as poly-generating HTFC [16]. Nonetheless, energy and emissions penalties associated with hydrogen transport must be accounted for in all analyses.

Cycle Configurations for Poly-Generating HTFC

Brouwer and Leal [21] investigated the production of hydrogen with high-temperature fuel cells by analyzing and comparing eight different cycle configurations using solid oxide fuel cells (SOFC) and molten carbonate fuel cells (MCFC). Six of the eight configurations use fuel cell heat to drive hydrogen production in an external reformer placed in different positions in the cycle. The other two configurations use the internal reformation capabilities of high-temperature fuel cells to produce hydrogen.

Based upon the cycle configurations developed by Brouwer and Leal [21], Margalef et al. [9] investigated in detail how placing the external reformer in different positions affects the fuel cell performance and the hydrogen production efficiency. In addition, Margalef et al. investigated the performance of the poly-generating plant at different fuel utilizations together with the integration of commercially available hydrogen separation and purification technologies (i.e., pressure swing adsorption) with the fuel cell balance of plant (BOP).

High-Temperature Fuel Cell Subsystem

High-temperature fuel cells (HTFC) generate electricity and heat through exothermic electrochemical reactions. Generated heat by the fuel cell reactions is typically utilized internally or externally by the endothermic fuel processing reactions, which in turn provide cooling to the system [22]. Surplus heat is used to preheat the fuel and oxidant streams before they enter the fuel cell and to produce the steam required for system operations. In addition, the remaining thermal energy contained in the exhaust gases can be used downstream of the fuel cell for poly-generation applications that require or value heat [22].

One possible configuration is to use the fuel cell heat to produce hydrogen via steam methane reforming in an external reformer (i.e., external reformation). Another possibility relies on the internal reforming capabilities of HTFCs and on the fact that the amount of high-quality heat produced by the exothermic reactions within the stack is typically greater than the heat required for fuel processing [6]. Therefore, more hydrocarbon fuel than that required for the electricity generation can be processed in an HTFC, creating a hydrogen-rich stream that could be subsequently purified and delivered at the point of production without the need of an external reformer [17] (i.e., internal reformation). This mode of operation implies lower stack fuel utilization factors and has been associated with synergies such as lower cell polarization losses and lower parasitic losses correlated with lower cooling air [16].

Separation/Purification Subsystem (Hydrogen Separation Unit (HSU))

Hydrogen separation and purification in refineries has been traditionally accomplished by using established technologies such as pressure swing adsorption (PSA), selective permeation processes using polymer membranes, or cryogenic separation process. Each process is based on a different separation principle, so each method differs significantly from each other. Economic aspects and other project considerations such as process flexibility, reliability, and scalability have to be taken into account to decide the hydrogen separation method [23].

Other hydrogen separation technologies include electrochemical hydrogen separation (EHS), which is foreseen as a promising technology to separate hydrogen fuel from a fuel cell anode exhaust stream. However, EHS technology is not currently mature and has not been used in poly-generating applications to date. In any case, the hydrogen-rich stream is treated and prepared in the hydrogen separation unit (HSU) according to the selected separation/purification technology. For this work, PSA technology has been selected due to its commercial readiness and recent use in poly-generating systems [24].

The HSU configuration depends upon the specific requirements of the hydrogen separation technology and the anode off-gas conditions. PSA technology requires relatively low inlet temperatures and high inlet pressures. Additionally, hydrogen separation with PSA becomes more efficient at high hydrogen partial pressures [23]. Therefore, to extract the hydrogen from the anode off-gas of a HTFC using a PSA, the HSU is required to:

- Decrease the hydrogen-rich stream temperature
- Increase the hydrogen-rich stream pressure
- Increase the hydrogen partial pressure of the hydrogen-rich stream

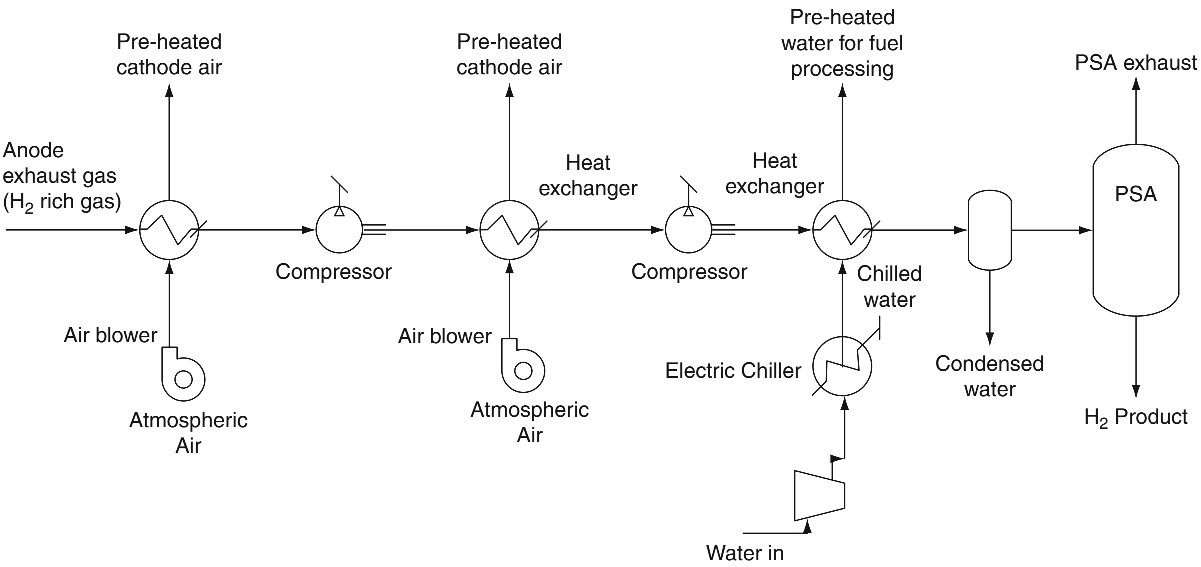
Table 1 shows representative PSA feed gas requirements and the current configuration design points.

A simplified HSU block consisting of a series of heat exchangers and compressors designed to meet the PSA temperature and pressure requirements is shown in Fig. 3.

As shown in Fig. 3, an electric chiller has been placed upstream of the PSA reactor to meet the temperature requirements when ambient air temperature (used as a cold media in the upstream heat exchangers) is too high. The electric chiller represents a small fraction of the total parasitic loads. Interestingly, required PSA inlet temperature is low enough to condense out sufficient water vapor from the gas stream. As shown in Fig. 3, condensed water may be removed upstream of the PSA reactor resulting

Hydrogen Production from High-Temperature Fuel Cells. Table 1 State-of-the-art PSA feed gas requirements

Parameter	State-of-the-art value range	Notes
Absolute pressure (kPa)	303–2,026	Based on the state of the art [25]
Temperature (°C)	4–50	Based on the state of the art [25]



Hydrogen Production from High-Temperature Fuel Cells. Figure 3
Hydrogen separation unit configuration

in higher hydrogen partial pressures and facilitating the PSA separation process.

In the internal reformation cases, the air and water that is heated in the HSU represents an opportunity to thermally integrate the HSU with fuel cell balance of plant. This thermal integration strategy is crucial in order to ensure overall thermal balance within the plant, which may be jeopardized at certain operating conditions if hydrogen is separated from the anode off-gas [16].

Integrated Poly-Generating HTFC Cycles According to Fuel Processing

Steam methane reformation (SMR) and water–gas shift (WGS) are the main fuel processing reactions that convert raw fuel (i.e., natural gas) into fuels more amenable to electrochemical oxidation (i.e., hydrogen and carbon monoxide) occurring on the nickel-based anode of HTFC.

Steam methane reforming is an endothermic reaction that consists of the reaction of methane and steam over a supported nickel catalyst to produce a mixture of hydrogen, carbon monoxide, carbon dioxide, and methane. The basic reforming reaction for methane is

$$CH_4 + H_2O \rightarrow 3H_2 + CO \quad \Delta H = 206\,\mathrm{kJ\,mol^{-1}} \tag{16}$$

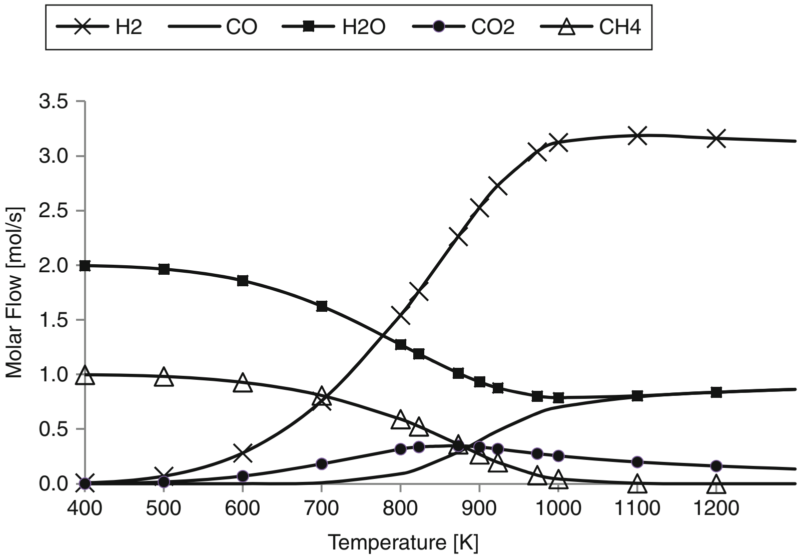
Heat has to be provided to drive the reaction forward to the hydrogen production direction. As shown in Fig. 4, hydrogen concentration under equilibrium conditions is highest between 900 and 1,100 K [17].

The water–gas-shift (WGS) reaction (starting from steam) is slightly exothermic and occurs at the same time as steam reforming [6]. During the shift reaction, additional hydrogen is produced. The basic water–gas-shift reaction for carbon monoxide is

$$CO + H_2O(g) \rightarrow H_2 + CO_2 \quad \Delta H = -41\,\mathrm{kJ\,mol^{-1}} \tag{17}$$

In poly-generating HTFC, if external reformer is used to produce hydrogen, special attention has to be given to the temperature at which heat from the fuel cell stack is transferred to the external reformer. Therefore, it is important to analyze different cycle configurations consisting on different positions of the external reformer within the system.

In the internal reformation cases where additional hydrogen production occurs at the same physical location than fuel cell reactions (i.e., direct internal reformation) or in a location that is thermally connected with the anode (i.e., indirect internal reformation), the temperature at which heat is



Hydrogen Production from High-Temperature Fuel Cells. Figure 4
Equilibrium composition as a function of temperature (steam-to-carbon ratio S/C = 2) [17]

transferred to the fuel processing reactions corresponds to the operating stack temperature which is equal to the temperature at which hydrogen production is maximized. This feature constitutes one of the key principles that lead to unprecedented high efficiencies achieved by poly-generating HTFC [8].

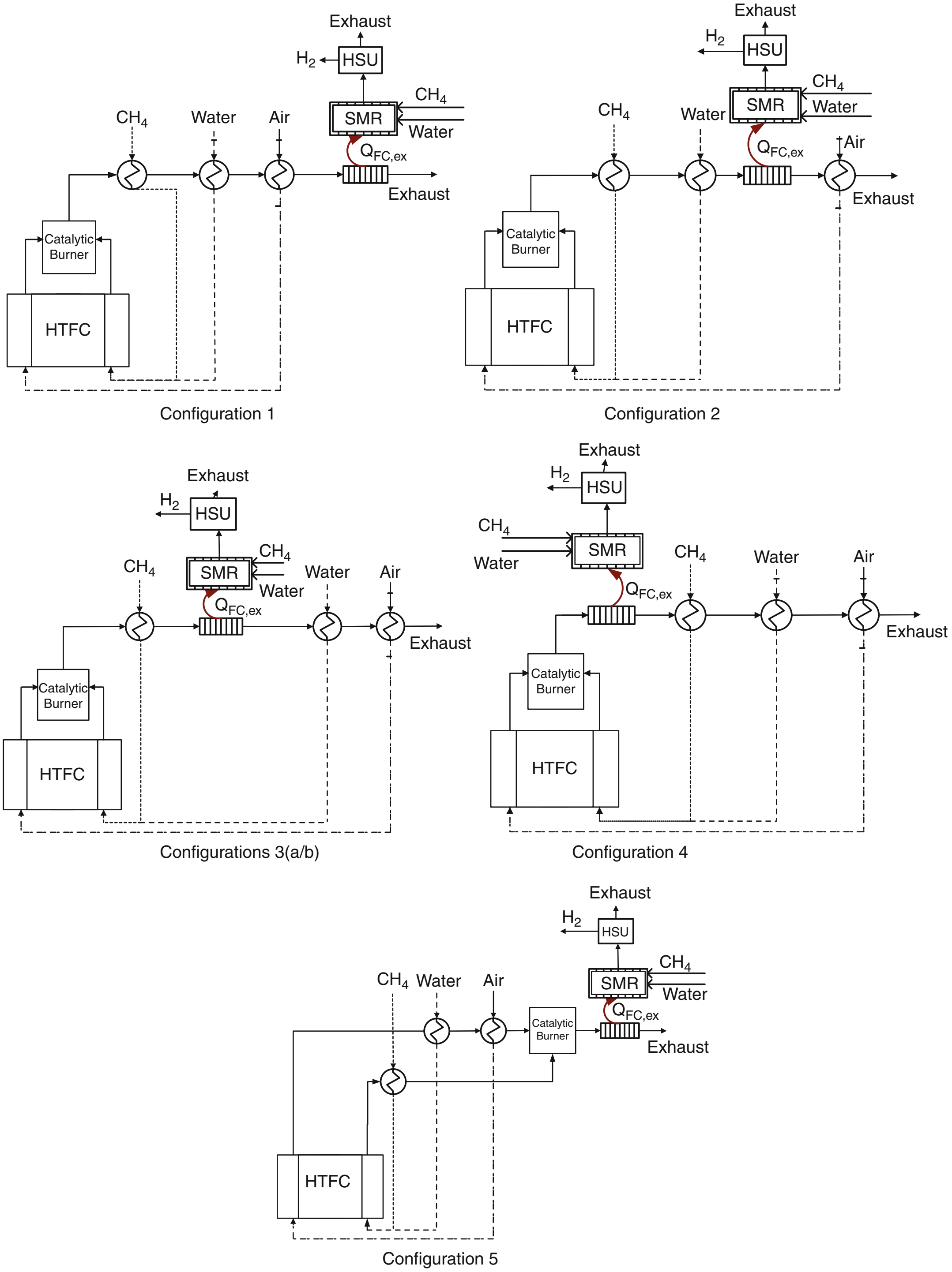
External Reformation Configurations High-temperature fuel cells (HTFC) can be integrated with an external reformer in multiple ways. Based on preliminary designs by Brouwer and Leal [17], Margalef et al. [9] analyzed in detail the external reforming configurations of poly-generating HTFC shown in Fig. 5.

In all cases, steam reformation is driven by the fuel cell exhaust heat in an external reformer, which takes as much heat as possible without compromising the fuel cell operating temperatures. Inlet temperatures of the fuel, steam, and air streams are kept constant at 1,173 K in order to sustain the electrochemical reactions within the stack. Therefore, depending upon the external reformer location, more or less heat is available to produce hydrogen with the external reformer. Note that in all the configurations, a hydrogen separation unit (HSU) block, based on PSA technology, is placed downstream of the reformer to separate and purify the hydrogen stream from the reformate gas.

Internal Reformation Configurations Internal reforming promotes hydrogen production within the fuel cell stack and provides cooling to the fuel cell stack due to its endothermic nature. Generally, fuel cell systems do not electrochemically consume all the fuel that is supplied (a fundamental limitation for all fuel cells) and they produce enough heat to reform much more fuel than the amount they consume. Remaining fuel exiting the anode presents a unique opportunity for low-cost hydrogen [17].

Margalef et al. [16] analyzed in detailed one internal reformation SOFC configuration at 80% and 60% fuel utilization (U_F). Figure 6 shows the schematic of this configuration. As shown, the HSU block is placed at the anode gas exit upstream the catalytic oxidizer.

Representative Performance Characteristics Modeling capabilities developed at the National Fuel Cell Research Center (NFCRC) of the University of California, Irvine have been used to evaluate the fuel cell performance and the hydrogen production capabilities of the presented poly-generating HTFC configurations. The complete model consists of an SOFC stack; heat exchangers to preheat the fuel, water, and air; an external SMR reactor placed in different locations for each of the different configurations; an adiabatic catalytic combustor that captures



Configuration 1

Configuration 2

Configurations 3(a/b)

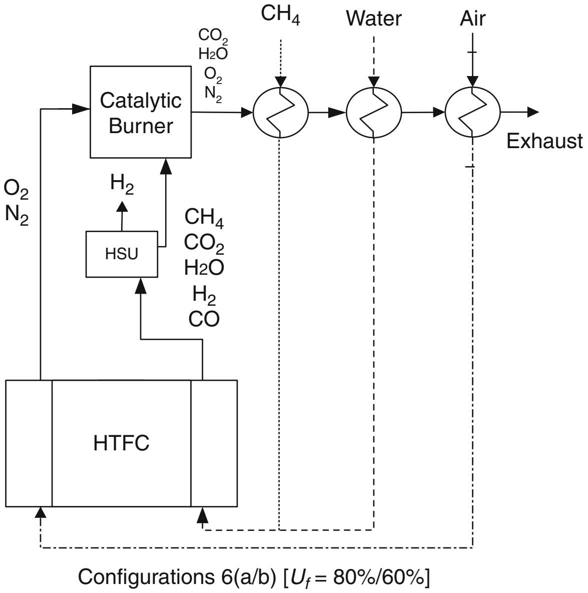
Configuration 4

Configuration 5

Hydrogen Production from High-Temperature Fuel Cells. Figure 5
Poly-generating HTFC configurations based upon external reformation

the thermal energy of the unused fuel downstream of the stack; and an HSU block based on PSA technology.

Each configuration has been analyzed following the same approach. For the external reformation cases (Conf. 1–5), stack input temperatures of the fuel, air, and steam have been fixed at 1,173 K whereas the amount of reformed methane varies depending upon how much heat is available after preheating all the input streams. Pinch analyses for each heat exchanger have been performed in order to avoid temperature crossovers within the heat exchangers. Similarly, for the internal reformation cases (Conf. 6a/6b), the amount of hydrogen extracted in the HSU block depends upon how much thermal energy has to be extracted from the anode off-gas in order to preheat the input streams to the specified temperatures [9].



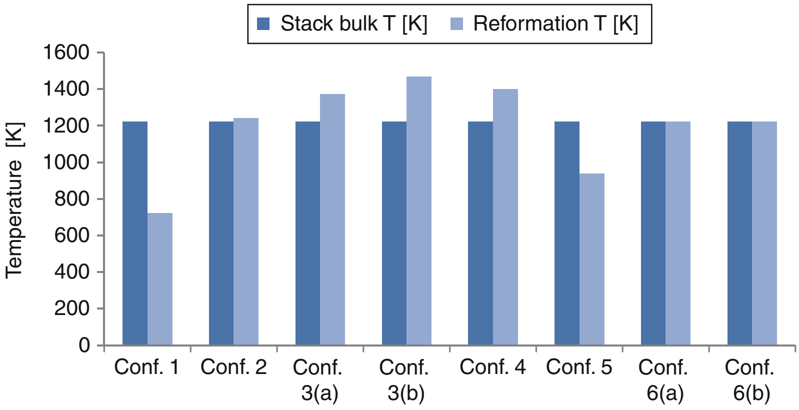
Hydrogen Production from High-Temperature Fuel Cells. Figure 6
Poly-generating HTFC based upon internal reformation

Bulk Stack and External Reformer Temperatures

For the external reformation cases, the heat available to produce hydrogen without compromising the thermal balance of the fuel cell determines how much methane will be taken by the external reformer. However, the hydrogen yield will be a function of the temperature at which the reformation takes places. As shown in Fig. 4, hydrogen yield peaks between 900 and 1,000 K and it flattens out after this point. Therefore, the external reformer should operate in this temperature range in order to maximize the hydrogen production.

Figure 7 shows both fuel cell stack and external reforming temperatures for all the configurations.

As expected, stack temperatures are the same for all cases since inlet stream temperatures have been fixed to a certain value. Importantly, in configurations 6a and 6b, stack and reformation temperatures are equal due to the fact that hydrogen is produced by internal reforming.



Hydrogen Production from High-Temperature Fuel Cells. Figure 7
Stack and reformer temperatures

Configuration 1 presents the lowest reformation temperature. This makes sense since the external reformer is placed downstream of all the heat exchangers where exhaust gas temperatures are lowest. Configuration 3b achieves the highest reformation temperature among the configurations. This configuration is the only external reforming configuration where the fuel utilization factor has been lowered to 60%. Therefore, more hydrogen will be oxidized in the catalytic combustor raising the exhaust gas temperature considerably. The rest of the configurations present similar reformation temperature values, all of them in range where hydrogen yield is high.

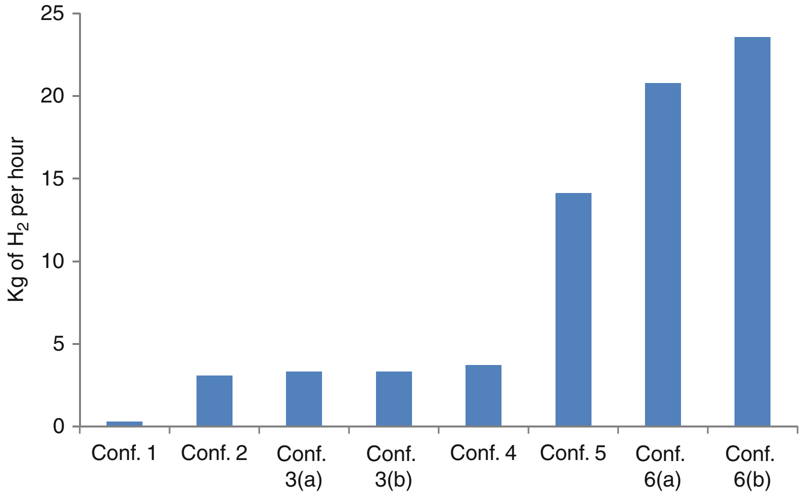
Hydrogen Production Rate Figure 8 shows the amount of hydrogen produced with each configuration. As observed, there is a significant difference between the hydrogen produced with the external reforming configurations and the amount of hydrogen produced with the internal reformation configurations.

The amount of hydrogen produced with configuration 1 is low. As shown in Fig. 7, the temperature at which reformation occurs in configuration 1 is 724 K, at which temperature hydrogen yield under equilibrium conditions is very small.

Figure 7 shows that in configurations 2, 3a, 3b, and 4, external reforming occurs at temperatures at which hydrogen yield is high. However, the amount of hydrogen produced is not comparable to the internal reformation cases or configuration 5. As mentioned previously, hydrogen production is a function of the temperature at which the reformation occurs but is also a function of the amount of transferred heat from the exhaust gas stream to the external reformer. Although in configurations 2, 3, and 4, steam reforming occurs at relatively elevated temperatures, the heat that can be transferred from the fuel cell exhaust to the reformer without compromising the fuel cell thermal balance is not enough to reform large amounts of methane. As a result, the amount of hydrogen that can be produced with these configurations is not large.

Configuration 5 presents higher hydrogen production than the other external reforming configurations. In this case, since the external reformer is placed after the catalytic combustor, the temperature at which the reformation occurs is high enough to achieve significant hydrogen yields. Importantly, since there is no preheater downstream of the reformer, more heat can be extracted from the fuel cell exhaust stream without affecting any fuel cell stream input temperatures.

Finally, configurations 6a and 6b achieve the highest hydrogen production. This is due to the fact that the reformation takes place within the SOFC stack, which operates within the range of temperatures



Hydrogen Production from High-Temperature Fuel Cells. Figure 8
Hydrogen production after PSA

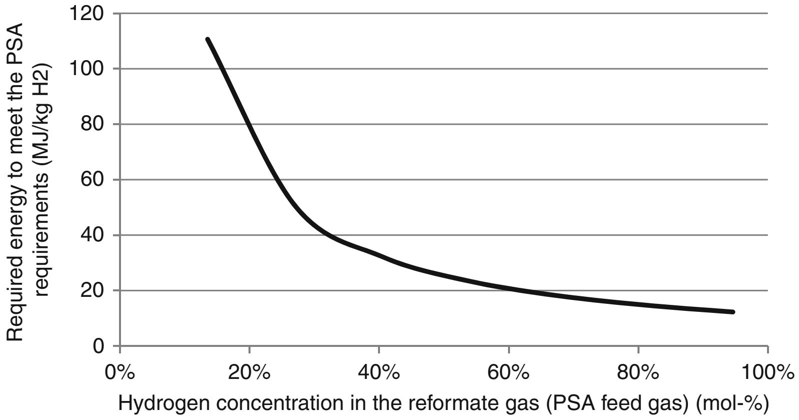
at which hydrogen yield is high. Additionally, the reformation reactions and the fuel cell reactions occur in the same physical space. Therefore, heat from the source (i.e., exothermic fuel cell reactions) to the sink (i.e., endothermic reformation reactions) is directly transferred without the need of a heat exchanger. Thus, more heat can be captured resulting in higher hydrogen yields. The total hydrogen production for each configuration is shown in Fig. 8.

Parasitic Loads Associated with the Hydrogen Separation Unit Energy required to separate the hydrogen from the reformate stream varies in each configuration. The HSU is required to increase the pressure and drop the temperature of the reformate gas in order to meet the PSA requirements. This process requires energy and as a result, decreases the overall efficiency of the poly-generating plant. Margalef et al. [9] estimated the amount of energy required to separate 1 kg of hydrogen with the HSU design presented in Fig. 3. The HSU model consists of a series of heat exchangers and compressors that decrease the temperature and increase the pressure of the reformate gas to the design point levels, accordingly, for each configuration. Figure 9 shows the energy per kilogram of hydrogen required for the preparation of the reformate gas to meet the PSA requirements, as a function of the molar concentration of hydrogen.

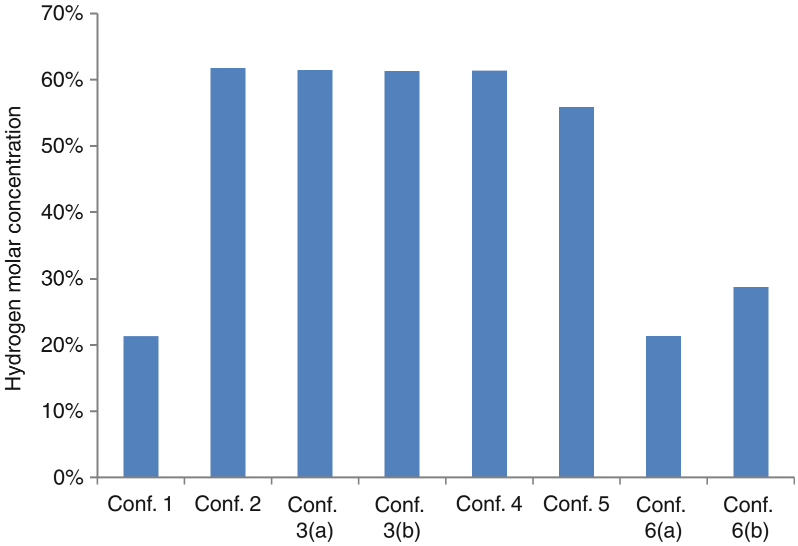
As seen, the energy required to separate 1 kg of hydrogen from the reformate stream does not decrease linearly with the hydrogen molar concentration and depends upon the hydrogen concentration in the anode off-gas. This is due to the fact that for lower hydrogen concentrations, relatively more gas has to be cooled and compressed to produce the same amount of hydrogen. Therefore, the amount of energy required to separate 1 kg of hydrogen with PSA technology is lower when the hydrogen concentration of the feed gas is high [9]. Figure 10 shows the molar hydrogen concentrations of the PSA feed gas for each configuration.

For all of the external reformation configurations except for configuration 1, hydrogen concentrations remain around 60%. Although the heat available to produce hydrogen in configuration 1 is comparable to the rest of the configurations, the temperature at which the reformation occurs is not high enough to achieve significant methane conversion. Similarly, configurations 6a and 6b present low hydrogen concentrations compared to the rest of the configurations. This makes sense since the produced hydrogen is mixed with the anode off-gas products which include all the carbon dioxide and steam products from the stack reactions, including internal reforming and electrochemical reactions. As a result, the hydrogen concentration of the reformate gas for the internal reformation cases is relative low when compared with the external reformation configurations.

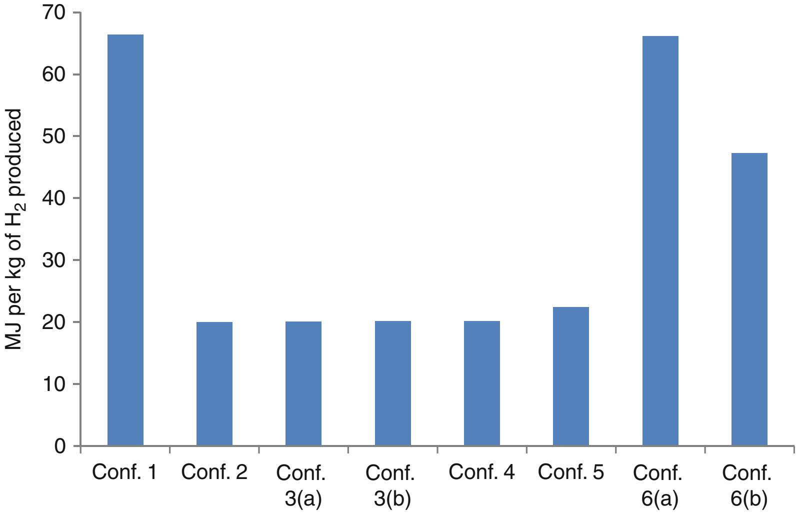
Figure 11 shows the energy required to separate 1 kg of hydrogen with each configuration. As shown, higher



Hydrogen Production from High-Temperature Fuel Cells. Figure 9
Specific energy required for the PSA process



Hydrogen Production from High-Temperature Fuel Cells. Figure 10
Hydrogen concentration in the PSA feed gas



Hydrogen Production from High-Temperature Fuel Cells. Figure 11
Parasitic load required to separate 1 kg of H_2 at the HSU

hydrogen concentration streams require less energy to separate the hydrogen from the reformate gas.

Comparative Efficiency Analyses

Because a poly-generating HTFC simultaneously produces electricity, hydrogen, and useful thermal energy, efficiencies can be measured and expressed in a number of different ways. Margalef et al. developed three different methods to appropriately calculate the overall and coproduct production efficiencies [8]. Table 2 shows the developed equations for each method. The methods have been labeled as: (1) *State-of-the-art Method*; (2) *Ideal Poly-generation Method*, and (3) *Supplemental Input Method*.

Hydrogen Production from High-Temperature Fuel Cells. Table 2 Efficiency equations for (1) state-of-the-art method, (2) ideal poly-generation method, and (3) supplemental input method [8]

		Electrical efficiency	Thermal efficiency	Hydrogen efficiency	Total mixed efficiency
1		$\frac{P_{net}}{E_{tot} - \frac{Q_{net}}{\eta_{boiler}} - \frac{H_2}{\eta_{SMR}}}$	$\frac{Q_{net}}{E_{tot} - \frac{P_{net}}{\eta_{CC}} - \frac{H_2}{\eta_{SMR}}}$	$\frac{H_2}{E_{tot} - \frac{P_{net}}{\eta_{CC}} - \frac{Q_{net}}{\eta_{boiler}}}$	$\frac{P_{net} + Q_{net} + H_2}{E_{tot}}$
2		$\frac{P_{net}}{E_{tot} + Q_{net} + H_2}$	$\frac{Q_{net}}{E_{tot} + P_{net} - H_2}$	$\frac{P_{PSA}}{E_{tot} - P_{net} - Q_{net}}$	$\frac{P_{net} + Q_{net} + H_2}{E_{tot}}$
3	External reforming	$\frac{P_{net}}{E_{tot} - \frac{Q_{e^-}}{\eta_{boiler}} - \left(F_{H_2} + \frac{P_{PSA}}{\eta_{CC}}\right)}$	$\frac{Q_{net}}{\frac{Q_{net}}{\eta_{boiler}}} = \eta_{boiler}$	$\frac{H_2}{F_{H_2} + \frac{P_{PSA}}{\eta_{CC}}}$	$\frac{P_{net} + Q_{net} + H_2}{E_{tot}}$
	Internal Reforming.	$\frac{P_{net}}{E_{tot} - \frac{Q_{e^-}}{\eta_{boiler}} - \left((U_f - U_{f,\mathrm{H2}ES})E_{tot} + \frac{P_{PSA}}{\eta_{CC}}\right)}$		$\frac{H_2}{(U_F - U_{f,\mathrm{H2}ES})E_{tot} + + \frac{P_{PSA}}{\eta_{CC}}}$	

Hydrogen Production from High-Temperature Fuel Cells. Table 3 Efficiency results obtained with the Supplemental Input Method (LHV)

Configuration	1	2	3a	3b	4	5	6a	6b
Electrical efficiency	53.4%	53.3%	53.3%	46.9%	53.3%	52.8%	50.0%	58.4%
Hydrogen efficiency	18.0%	73.5%	73.2%	68.2%	73.2%	62.1%	90.7%	83.5%
Overall efficiency	52.3%	54.7%	54.8%	48.5%	55.0%	55.7%	70.0%	69.5%

These methods are based upon different and reasonable assumptions for the allocation of input energy to each of the coproducts. It should be clear that there is not a unique or superior method for calculating poly-generation efficiencies so that each of the methodologies proposed can be used in comparative analyses if based upon accurate assumptions. For the current specific analyses, the *Supplemental Input Method* is used to estimate the overall electrical and hydrogen production efficiencies for each configuration. Thermal efficiency has not been calculated since it depends upon the thermal requirements of each specific application and there may be cases when thermal energy is not required. Results are presented in Table 3.

The efficiency results shown in Table 3 correspond to the Supplemental Input Method [8] and do not include the heat products shown in the equations presented in Table 2. With this method, electrical efficiency is the net power output P_{net} divided by the energy flow allocated exclusively for the electricity production, which corresponds to the total energy flow in E_{tot} minus the energy flow that has been specifically used to produce hydrogen product. The energy flow used to produce hydrogen product includes feedstock energy (i.e., additional fuel) as well as the necessary fuel to generate the electricity required for the hydrogen separation as if it was produced with a state-of-the-art combined cycle plant. The lowest electrical efficiency value corresponds to configuration 3b. In this case, hydrogen is produced externally and the fuel cell is operating at 60% utilization factor. As expected, the electrical efficiency is low since more fuel is used without obtaining any additional energy flow output (i.e., hydrogen fuel). On the other hand, the highest electrical efficiency value corresponds to configuration 6b, in which hydrogen is produced internally and the fuel cell is operated at 60% utilization factor.

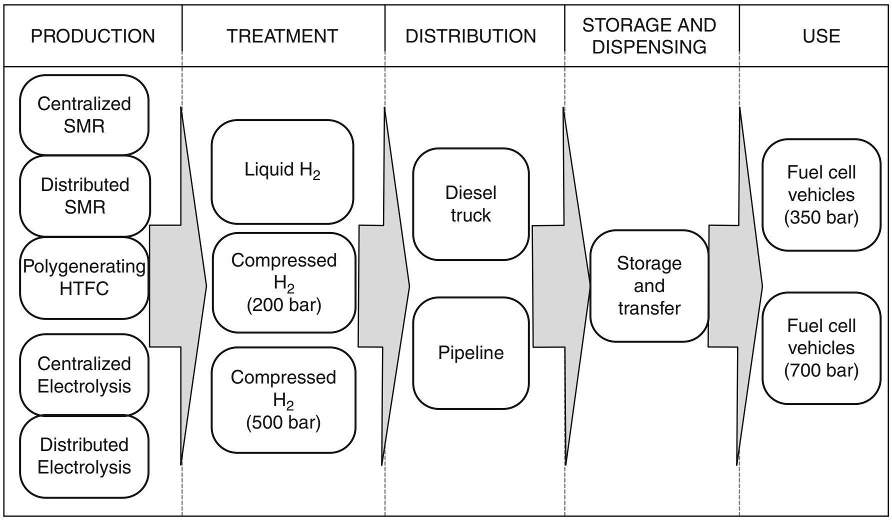
Due to the synergies associated with higher voltages at lower fuel utilizations, electrical efficiency is significantly higher than the rest of the configurations. Nevertheless, one must recall the discussion associated with the key principles above that this analysis confirms that these synergies are only captured if hydrogen is separated from the anode off-gas. Otherwise, more fuel is being used to obtain the same mix of energy outputs. The rest of configurations present similar electrical efficiencies.

With the Supplemental Input Method, hydrogen efficiency is calculated in a similar way as specified in Table 2. The chemical power output of the hydrogen produced, H, is divided by the energy flow input specifically allocated to produce hydrogen. Once again, it corresponds to the feedstock energy (i.e., additional fuel) as well as the necessary fuel to generate the electricity required for the hydrogen separation as if it was produced with a state-of-the-art combined cycle plant. As seen in Table 3, the highest values correspond to the internal reformation cases. Interestingly, although the parasitic load per kilogram of hydrogen produced is higher in configuration 6a than in configuration 6b (see Fig. 11), hydrogen production efficiency is greater in the former case due to the fact that when hydrogen is produced at 80% utilization factor, the additional fuel feedstock allocated to produce hydrogen is equal to zero.

Finally, overall efficiency values are similar in all the external reformation cases, even for configuration 1 where the hydrogen output is almost negligible. This indicates that the hydrogen production does not affect the overall performance when the amount of hydrogen is relatively small (i.e., external reformation cases). As expected, both internal reformation cases achieve the highest overall fuel-to-product efficiency values.

Hydrogen Transport Impacts on Comparative Analyses

Today, most hydrogen is produced from natural gas in large steam-methane reformation (SMR) plants in a centralized fashion. To produce hydrogen at large scale in a centralized plant is more efficient than producing hydrogen at small scale [26]. Stephens-Romero et al. [20] define production, treatment, distribution, storage, and dispensing as the main steps of a well-to-tank hydrogen supply chain. Each step can be accomplished with a variety of different technologies. Figure 12 shows the main steps of a generic hydrogen supply chain from well to tank with the most widely used technologies.



Hydrogen Production from High-Temperature Fuel Cells. Figure 12
Steps in well-to-tank efficiency analysis of H_2

One of the main benefits of producing hydrogen in a distributed manner versus centralized manner is the elimination of the fuel transport and delivery steps (fuel distribution) which can be energy intensive. Furthermore, if hydrogen is produced on demand as in poly-generating HTFC, hydrogen storage may be considerably reduced compared with the distributed reformer case.

Comparative Analysis of Ten Different Hydrogen Supply Chains

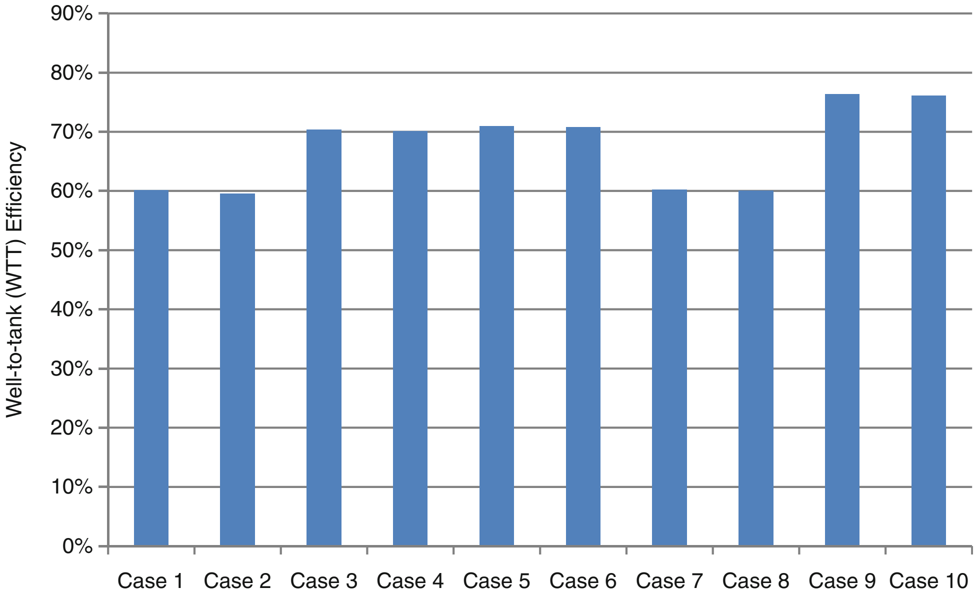
To evaluate the benefits of producing hydrogen in a distributed fashion versus centralized production, different hydrogen supply chains with different technologies for each step are analyzed from an energy perspective, resulting in ten different cases. Details of each supply chain are provided in Table 4. As shown, two different final uses corresponding to the two available fuel cell vehicle tank pressures (i.e., 350 bar and 700 bar) [27] are evaluated. In addition, when hydrogen is produced in a centralized SMR plant, the three available treatment methodologies are investigated. As a result, ten different supply chains have been evaluated.

After evaluating the energy required for each step from *production* to *dispensing* of hydrogen, well-to-tank (WTT) efficiencies of the ten different hydrogen supply-chain scenarios were calculated. Figure 13 shows the main results obtained from the analyses.

As shown, the lowest WTT efficiency values are observed for the distributed steam- methane reformation (SMR) cases whereas the highest values correspond to poly-generating HTFC. As expected, due to the significant energy penalties associated with liquefaction processes, WTT efficiencies of centralized SMR with liquid hydrogen transportation are almost as low as the distributed SMR cases. Importantly, the production step includes the energy content of the fuel feedstock. The WTT efficiency values have

Hydrogen Production from High-Temperature Fuel Cells. Table 4 Definition of the ten unique hydrogen supply chain cases

Case	Production	Treatment	Distribution	Storage	Dispensing
1	Centralized SMR	Liquid H_2	Diesel truck (100 km)	Liquid H_2	Use = 350 bar (Transfer = 7 bar to 400 bar)
2	Centralized SMR	Liquid H_2	Diesel truck (100 km)	Liquid H_2	Use = 700 bar (Transfer = 7 bar to 800 bar)
3	Centralized SMR	Compressed H_2 (200 bar)	Diesel truck (100 km)	Compressed H_2 (200 bar)	Use = 350 bar (Transfer = 200 bar to 400 bar)
4	Centralized SMR	Compressed H_2 (200 bar)	Diesel truck (100 km)	Compressed H_2 (200 bar)	Use = 700 bar (Transfer = 200 bar to 800 bar)
5	Centralized SMR	Compressed H_2 (500 bar)	Diesel truck (100 km)	Compressed H_2 (500 bar)	Use = 350 bar (Transfer = 500 bar to 400 bar)
6	Centralized SMR	Compressed H_2 (500 bar)	Diesel truck (100 km)	Compressed H_2 (500 bar)	Use = 700 bar (Transfer = 500 bar to 800 bar)
7	Distributed SMR	Compressed H_2 (500 bar)	n/a (0 km)	Compressed H_2 (500 bar)	Use = 350 bar (Transfer = 500 bar to 500 bar)
8	Distributed SMR	Compressed H_2 (500 bar)	n/a (0 km)	Compressed H_2 (500 bar)	Use = 700 bar (Transfer = 500 bar to 800 bar)
9	Poly-generating HTFC	Compressed H_2 (500 bar)	n/a (0 km)	Compressed H_2 (500 bar)	Use = 350 bar (Transfer = 500 bar to 500 bar)
10	Poly-generating HTFC	Compressed H_2 (500 bar)	(0 km)	Compressed H_2 (500 bar)	Use = 700 bar (Transfer = 500 bar to 800 bar)



Hydrogen Production from High-Temperature Fuel Cells. Figure 13
Well-to-tank efficiencies of the evaluated supply-chains

been obtained by dividing the energy content of 1 kg of hydrogen (on a LHV basis) by the total energy required from WTT corresponding to each case.

From these energy-based comparative analyses, it is concluded that, of all the analyzed cases, poly-generating HTFC provide the most efficient way to produce and deliver hydrogen.

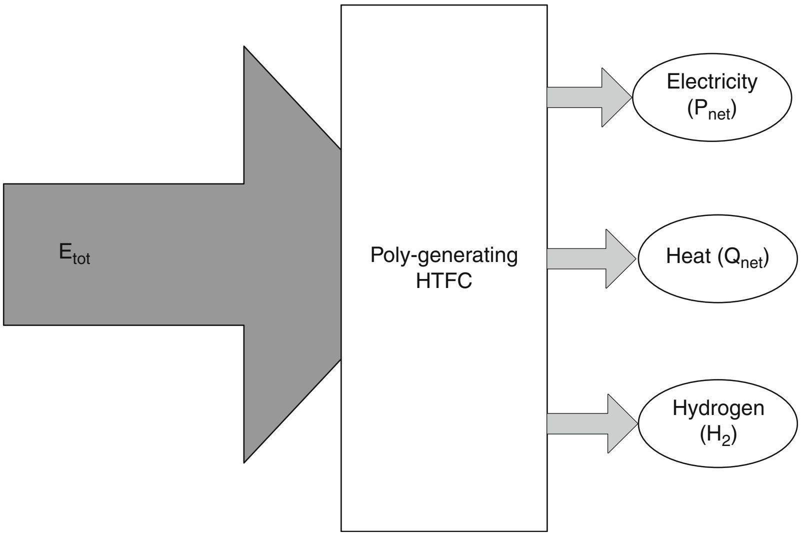
Discussion of Scientific Basis for Observed Performance

High-temperature fuel cells (HTFC) operate at high enough temperature and produce enough heat and water from the electrochemical reactions to provide heat to the internal fuel processing reactions (i.e., internal reformation). Importantly, the overall heat generated by the fuel cell reactions is typically greater than the heat consumed by the endothermic fuel processing reactions occurring within the fuel cell stack [6]. Surplus heat is partially used to preheat the fuel and oxidant streams before they enter the fuel cell and to produce the steam required for system operations. Therefore, more hydrocarbon fuel than that required for the electricity generation could be processed in an HTFC, creating a hydrogen-rich stream that could be subsequently purified and delivered to the point of use without the need of an external reformer. This represents the main concept that defines and allows the production of hydrogen in addition to electricity and heat with an HTFC. Such systems have been identified as poly-generating HTFC or tri-generation systems (Fig. 14).

Poly-generation systems can produce and deliver hydrogen fuel with lower marginal costs, fuel use, and emissions than conventional hydrogen production and delivery methods [28]. Since hydrogen is produced in a distributed generation fashion there is no energy or emissions penalty associated with the transport and delivery of the hydrogen to its point of use. Recent studies performed at the National Fuel Cell Research Center (NFCRC) of the University of California Irvine demonstrate that less energy is needed to transport and deliver hydrogen to vehicles compared with distributed and centralized steam methane reforming (SMR) resulting in higher well-to-tank (WTT) efficiencies [16].

Definition and Quantification of the Synergies Associated with Poly-Generating HTFC

Among the main benefits associated with the production of hydrogen with HTFC, there are synergistic



Hydrogen Production from High-Temperature Fuel Cells. Figure 14
Inputs and outputs of a poly-generating HTFC

effects intimately related with operating the fuel cell at low fuel utilization factors:

1. Synergistic impact of lower fuel utilization leads to higher cell voltages which result in:
   (a) Higher electrochemical efficiencies
   (b) Less electrochemical heat generated per mol of input fuel
2. Synergistic impact of lower fuel utilization increases the heat sink associated with the endothermic SMR reaction (i.e., internal fuel processing), reducing excess air requirements for cooling.

Synergies 1(b) and 2 reduce the excess air requirements for the stack cooling. Consequently, lower cooling requirements will lower the parasitic loads from the cathodic air blowers, increasing the net power of the system and the overall system efficiency. In this section, more details about each of the identified synergies are provided.

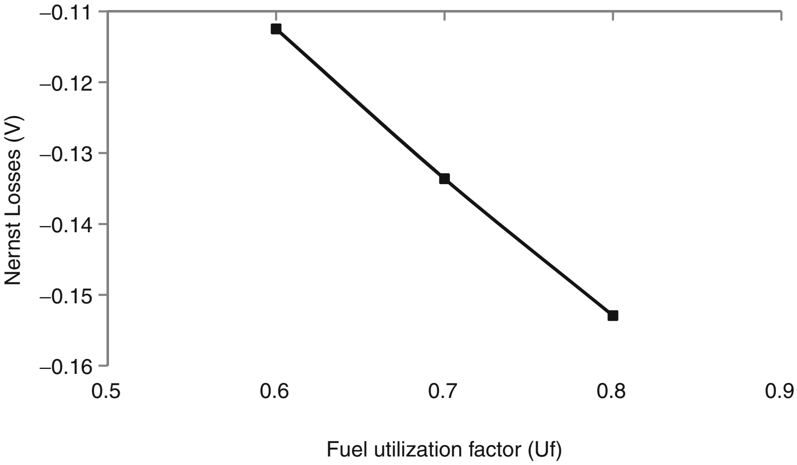
Synergistic Impact of Lower Fuel Utilization Leads to Higher Electrochemical Efficiencies Ideal reversible voltage (E_0) drops as soon as current is being produced due to the changes in product and reactant concentrations. Importantly, reactants and products concentrations along the electrode channels vary at different utilization factors. Therefore, operating the fuel cell at different fuel utilizations may affect the cell voltage. The voltage reduction caused by the change of species concentrations is known as *Nernstian losses* (η_{Nernst}), which added to the reversible voltage E_0, results on the Nernst Voltage (V_{Nernst}).

$$\frac{V_{Nernst} = E_0 + \eta_{Nernst} = E_0 + \frac{RT}{2F} In\left(x_{H_2} x^{\frac{1}{2}}_{o_2}\right)}{xH_2O} \tag{18}$$

Figure 15 shows the Nernstian losses (η_{Nernst}) as a function of the utilization factor obtained with the poly-generating SOFC modeling efforts by Li and Margalef [16, 30].

As shown in Fig. 15, voltage drop related with the Nernstian losses decreases at lower fuel utilizations. This effect represents in fact one of the synergies associated with poly-generating fuel cells, which typically operate at lower fuel utilizations than stand-alone fuel cells (i.e., electricity and heat production only).

To quantify this synergy, the system can be analyzed at two different fuel utilizations. The first value corresponds to the typical utilization factor of a stand-alone SOFC (e.g., $U_f = 0.9$) whereas the second value corresponds to the typical utilization factor of a poly-generating SOFC (e.g., $U_f = 0.6$). The difference in Nernst losses (V_{Nernst}) between both operating points represents the voltage gain of operating the fuel cell at lower fuel utilizations and quantifies this synergistic effect. Approximately, a gain of 0.06 V in the



Hydrogen Production from High-Temperature Fuel Cells. Figure 15
Nernst losses (η_{Nernst}) versus the fuel utilization factor

Nernst voltage has been observed by operating the fuel cell at typical utilization factor used in poly-generating fuel cells ($U_f = 0.6$ compared to $U_f = 0.9$).

Additionally, the utilization factor affects the fuel cell performance through reaction kinetics since they also depend upon the reactant and product concentrations at the reaction sites [22]. The voltage drop associated with the reaction kinetics at different fuel utilizations is known as *reaction loss* ($\eta_{rxn\ \mathrm{rate}}$). However, previous analyses performed by Margalef et al. [16] show that reaction losses represent a small fraction of the total voltage drop related with the concentration changes associated with different fuel utilizations. Therefore, the overall effects of operating the fuel cell at lower fuel utilizations on the cell voltage can be defined as

$$\Delta V_{\mathrm{cell,conc}} = \Delta V_{\mathrm{cell,Nernstian}} + \Delta V_{\mathrm{cell},rxn\ \mathrm{rate}} \sim \Delta V_{\mathrm{cell,Nernstian}} \quad (19)$$

If all the energy from the hydrogen fuel were transformed into electrical energy, the cell voltage would correspond to the maximum voltage $v_{\max}$. The maximum voltage would be given by

$$V_{\max} = \frac{-\Delta \bar{g}_f}{n_i F} = 0.97\ V \text{ if using the LHV} \quad (20)$$

where $\Delta \bar{g}_f$ is the Gibbs free energy at 900°C, n_i is the number of electrons transferred during the electrochemical oxidization of fuel species i ($n_i = 2$ for H_2 and CO, $n_i = 8$ for CH_4), and F is the Faraday's constant. This voltage would be obtained from a 100% efficient system [6]. Therefore, the actual efficiency of the cell (or electrochemical efficiency) is the actual voltage divided by the maximum voltage, or

$$\eta_{\mathrm{cell}} = \frac{V_{\mathrm{cell}}}{V_{\max}} = \frac{V_{\mathrm{cell}}}{0.97} \text{ if using the LHV} \quad (21)$$

Therefore, the electrochemical efficiency gain associated with lower utilization factors can be written as

$$\Delta \eta_{\mathrm{cell}} = \frac{\Delta V_{\mathrm{cell,conc}}}{V_{\max}} = \frac{0.06}{0.97} \times 100 = 6.2\% \text{ if using the LHV} \quad (22)$$

This result demonstrates and quantifies the first synergy associated with generating hydrogen with HTFC. As seen, by lowering the utilization factor at typical values of poly-generating SOFC, the electrochemical efficiency increases by 6.2%.

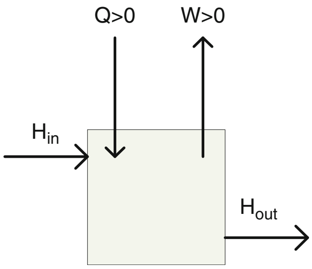
A more philosophical approach to explain the cell voltage drop as the fuel utilization factor increases is related to the mixing of products with reactants to dilute electrochemically active species concentrations. Water product mixes with the anode gas as it is formed throughout the electrochemical reactions. Note that mixing of products with reactants can occur in either the anode (e.g., MCFC, SOFC) or cathode (e.g., PEMFC, PAFC) compartments. The concentration of water product in the anode increases as the fuel utilization factor increases. Mixing processes are irreversible so they produce entropy. As a result, a completely reversible SOFC operation would be only approached as $U_f \rightarrow 0$. In other words, irreversibilities increase as the utilization factor increases. Consequently, one wants to operate a fuel cell at the lowest possible utilization factor in order to reduce the mixing irreversibilities. However, there is a compromise between the cell efficiency and the fuel utilization efficiency that counterbalances this consideration with a desire for high utilization to achieve high overall system electrical efficiency.

Synergistic Impact of Lower Fuel Utilization Increases the Heat Sink Associated with the Endothermic SMR Reaction, Reducing Excess Air Requirements for Cooling A high-temperature fuel cell (HTFC) is an electrochemical device that converts chemical energy of a fuel and oxidant directly to electricity. Fuel cell electrochemical reactions are exothermic, thus electrochemical heat is produced. In addition, during the operation of an HTFC, ohmic resistance across the electrodes and bipolar plates generates heat. The specific electrochemical heat released qFC can be calculated applying the first law of thermodynamics, according to Eq. 23. Sign convention for this analysis is specified in Fig. 16:

$$\Delta\bar{h}_f = qFC - wFC \tag{23}$$

where $\Delta\bar{h}_f$ is the enthalpy of combustion of hydrogen corresponding to the reaction

$$H_2 + \frac{1}{2}O_2 \rightarrow H_2O(\text{liquid}) \quad \Delta\bar{h}_f = -285.84\,\text{kJ}\,\text{mol}^{-1} \tag{24}$$



Hydrogen Production from High-Temperature Fuel Cells. Figure 16
Sign convention

and w_{FC} is the electric work produced by the fuel cell, given by the Eq. 23

$$w_{FC} = n_e F V_{\text{cell}} \tag{25}$$

where V_{cell} is the operational cell voltage, n_e is the number of electrons transferred during the electrochemical reactions (i.e., $n_e = 2$ for hydrogen), and F is the Faraday's constant. Combining Eq. 23 and Eq. 25, the specific heat generated during a real fuel cell operation can be calculated as follows:

$$qFC = \Delta\bar{h}_f + w_{FC} = \Delta\bar{h}_f + n_e F V_{\text{cell}} \left[\frac{\text{kJ}}{\text{mol}\,H_2}\right] \tag{26}$$

In a real fuel cell, although part of the heat is exchanged and used in several ways, including fuel processing and air preheating, excess air is usually required to prevent overheating [29]. Assuming that the fuel processing consists of steam reformation (SMR) and water–gas shift (WGS), the total specific heat absorbed by the fuel processing reactions qFC including the steam production is

$$\begin{aligned} qFP &= \Delta\bar{h}_{SMR} + \Delta\bar{h}_{WGS} + 2 \times \Delta\bar{h}_{vap} \\ &= 206.1 - 41.2 + 2 \times 44.1 = 253.1 \left[\frac{\text{kJ}}{\text{mol}\,CH_4}\right] \end{aligned} \tag{27}$$

where $\Delta\bar{h}_{SMR}$ is the enthalpy of steam reforming reaction, $\Delta\bar{h}_{WGS}$ is the enthalpy of water–gas-shift reaction, and $\Delta\bar{h}_{vap}$ is the enthalpy of vaporization of water. Therefore, remaining heat that has to be transferred to the excess air for cooling purposes q_{cool} is the difference between the heat liberated during the fuel cell reactions qFC and the heat absorbed by the fuel processing reactions qFP:

$$q_{\text{cool}} = m \cdot qFC + qFP \left[\frac{\text{kJ}}{\text{mol}\,CH_4}\right] \tag{28}$$

where m is the number of moles of hydrogen produced during the fuel processing reactions. For this specific case, the overall fuel processing reaction includes SMR and WGS, and is given by Eq. 29.

$$CH_4 + 2H_2O \rightarrow 4H_2 + CO_2 \tag{29}$$

Therefore, m is equal to 4. Combining Eqs. 27–29, the specific cooling required can be written as

$$q_{\mathrm{cool}} = 4\left(\Delta\bar{h}_f + n_e F V_{\mathrm{cell}}\right) + qFP\left[\frac{\mathrm{kJ}}{\text{mol of CH}_4}\right] \quad (30)$$

As expected, more electrochemical heat is liberated per mole of hydrogen qFC at higher fuel utilizations. Consequently, by operating the fuel cell at lower utilization factors reduces the fuel cell cooling requirements per mole of input fuel, due to the higher cell voltages that result in less electrochemical heat generated. Lower cooling requirements mean lower parasitic loads from the air blowers, increasing the net power of the system and the overall system efficiency.

World's First Poly-Generating HTFC System

The world's first installation to poly-generate electricity and hydrogen with high-temperature fuel cells (i.e., a molten carbonate fuel cell) has been installed at the Orange County Sanitation District (OCSD) in Fountain Valley, California [24]. The system that has been installed at OCSD is the world's first hydrogen coproduction system installed at a customer site. In addition, it has begun to produce electricity, heat, and hydrogen from renewable digester gas fuel. The prime contractor in the effort is Air Products and Chemicals, Inc., who is working with FuelCell Energy and the National Fuel Cell Research Center and OCSD on the project. The main funding agencies are the US Department of Energy, California Air Resources Board, and the South Coast Air Quality Management District.

This important installation of a prototype poly-generating HTFC operating on digester gas at OCSD will provide important data and experience that will benefit the advancement of the poly-generation concept. The installation itself and measured performance characteristics may also identify important areas for research, development, improvement, and application of the concept.

Future Directions

Recommendations for future research work on poly-generating HTFC development include the development of a detailed dimensional and dynamic SOFC model that is required to solve the complicated internal temperature and species concentration profiles that are not resolved by nondimensional models. Thermodynamic and nondimensional SOFC insights are nonetheless useful for system design and comparative analyses. In addition, the overall system performance obtained by these models generally sets the ideal system performance target. However, since the outlet flow temperatures do not necessarily represent the maximum temperatures in the SOFC, and peak temperatures often occur internal to the stack, a dimensional SOFC model and/or internal temperature measurements could be used to determine if the SOFC working voltage is achievable and practical.

All the analyses performed in previous literature have been performed at a fixed current density value. Therefore, the same analyses described in this chapter should be performed at lower current densities. At such operating conditions, it is likely that the operational cell voltage will rise, increasing the electrochemical efficiency. However, less electrochemical heat will be generated, which is likely to reduce the hydrogen production potential. Therefore, it is possible to adjust the system performance for each specific market depending upon which product is more valuable by varying the operational current density in addition to the fuel utilization.

It has been found that pressure swing adsorption (PSA) might not be the best technology for extraction and purification of the hydrogen from the anode off-gas due to the PSA feed gas requirements. A complex HSU configuration is required in order to meet pressure, temperature, and hydrogen concentration requirements. Therefore, different purification strategies such as electrochemical hydrogen separation and selective permeation membranes might be more suitable for the poly-generation application.

An economic analysis of poly-generating high-temperature fuel cells should be performed in detail. From an efficiency perspective and considering only the configurations presented herein, poly-generating HTFC are the best solution. However, the economics of such novel systems should be analyzed and compared with conventional production methods. Additionally, due to the poly-generating nature of these systems, different strategies to fit each of the specific market requirements should be designed. These strategies might be resolved by dynamic modeling efforts to investigate the ability of system configurations to dynamically dispatch each of the respective products depending upon demand profiles for each product.

Bibliography

1. Samuelsen GS (2005) Energy station concept. In: California hydrogen blueprint plan, vol 2. California Environmental Protection Agency, Sacramento. http://www.hydrogenhighway.ca.gov/plan/reports/volume2_050505.pdf
2. EG & G Technical Services, Inc. (2004) Fuel cell handbook, 7th edn. U. S. Department of Energy Office of Fossil Energy, Morgantown
3. Brouwer J, Jabbari F, Leal EM, Orr T (2006) Analysis of a molten carbonate fuel cell: numerical modeling and experimental validation. J Power Sour 158(1):213–224
4. Sundmacher K et al (2007) Molten carbonate fuel cells modeling, analysis, simulation, and control. Wiley, Weinheim
5. Aguiar P, Chadwick D, Kershenbaum L (2002) Modelling of an indirect internal reforming solid oxide fuel cell. Chem Eng Sci 57:1665–1677
6. Larminie J, Dicks A (2003) Fuel cell systems explained. Wiley, West Sussex
7. Online resource available at: http://www.energy.ca.gov/distgen/
8. Margalef P, Brown T, Brouwer J, Samuelsen GS (2010) Efficiency of polygenerating high temperature fuel cells. J Power Sour POWER_POWER-D-10-02363
9. Margalef P, Brown T, Brouwer J, Samuelsen GS (2010) Conceptual design and configuration performance analyses of polygenerating high temperature fuel cells. Intl J Hydrog Energy. Accepted 11 May 2011
10. Dufour J et al (2009) Life cycle assessment of processes for hydrogen production. Environmental feasibility and reduction of greenhouse gases emissions. Intl J Hydrog Energy 34:1370–1376
11. Holladay JD, Hu J, King DL, Wang Y (2009) An overview of hydrogen production technologies. Catal Today 139(2009):244–260
12. California Fuel Cell Partnership Action Plan (2009) Hydrogen fuel cell vehicle and station deployment plan: a strategy for meeting the challenge ahead. California Fuel Cell Partnership Action Plan, West Sacramento
13. Spath P, Mann M (2001) Life cycle assessment of hydrogen production via natural gas steam reforming. NREL/TP-570-27637. NREL National Renewable Energy Laboratories, Golden
14. Ogden JM (2002) Review of small stationary reformers for hydrogen production. Technical Report to the International Energy Agency
15. Sverdrup GM et al (2006) Status of hydrogen production pathways – comparison of energy efficiencies, fossil fuel use, greenhouse gas emissions, and costs. In: WHEC 16, 13–16 June 2006, Lyon
16. Margalef P (2010) On the polygeneration of electricity, heat and hydrogen with high temperature fuel cells. PhD dissertation, University of California, Irvine
17. Brouwer J, Leal E (2006) A thermodynamic analysis of electricity and hydrogen polygeneration using a solid oxide fuel cell. J Fuel Cell Sci Technol 3:137–143
18. Singhal SC, Kendall K (2003) High temperature solid oxide fuel cells: fundamentals, design and applications. Elsevier, Oxford
19. Cengel YA, Boles MA (2006) Thermodynamics: an engineering approach, 6th edn. McGraw Hill, Boston
20. Stephens-Romero S (2010) Systematic planning to optimize investments in hydrogen infrastructure deployment. Intl J Hydrog Energy 35:4652–4667
21. Brouwer J, Leal E (2005) Production of hydrogen using high-temperature fuel cell: energy and exergy analysis. In: 18th international congress of mechanical engineering. ABCM, Ouro Preto
22. O'Hayre R, Cha S, Colella W, Prinz FB (2006) Fuel cell fundamentals. Wiley, New York
23. Miller QM, Stocker J (1989) Selection of a hydrogen separation process. In: NPRA Annual Meeting held 19–21 March 1989, San Francisco
24. Heydorn EC, Patel P (2010) Development of a renewable hydrogen station. ICEPAG, Costa Mesa
25. QuestAir H-3200 Commercial Brochure. http://www.xebecinc.com/hydrogen-purification-h3200.php
26. Bossel U (2003) Energy and the hydrogen economy
27. Air Products & Chemicals Inc. Recovery process using pressure swing adsorption technology. http://texasiof.ces.utexas.edu/texasshowcase/pdfs/presentations/c6/pcook.pdf
28. Vollmar HE et al (2000) Innovative concepts for the coproduction of electricity and syngas with solid oxide fuel cells. J Power Sour 86:90–97
29. Van Herle J et al (2003) Energy balance model of a SOFC cogenerator operated with biogas. J Power Sour 118:375–383
30. Li M (2010) Detailed fuel cell modeling for coal-based integrated gasification fuel cell system design and analysis. Mechanical and aerospace doctoral studies dissertation, University of California, Irvine

Hydrogen Production Science and Technology

TIMOTHY E. LIPMAN
Institute of Transportation Studies, University of
California – Berkeley, Berkeley, CA, USA

Article Outline

Introduction
Review of Hydrogen Production Methods
Reformation of Fossil Hydrocarbons and Bioderived Liquids
Hydrogen Production Costs
Conclusion
Bibliography

Introduction

Hydrogen is an abundant element that is currently produced for a wide range of industrial uses. In recent years, it is also starting to be used for stationary power production in fuel cell systems and to power fuel cell electric vehicles including buses, forklifts, and passenger cars. One highly compelling attribute of hydrogen is that it can be produced in many different ways and using a variety of feedstocks; however, it does have to be *produced* through one of a variety of processes rather than simply "discovered" and delivered to market as is, for example, natural gas.

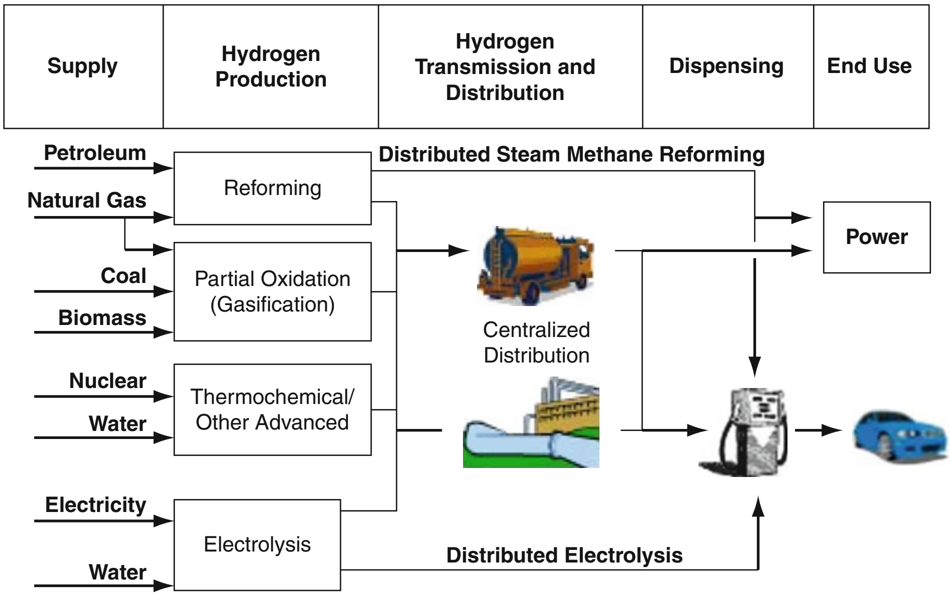
Approximately 10–11 million metric tonnes of hydrogen is produced in the USA each year [1]. For reference, this is enough to power 20–30 million cars (using 700–1,000 gal energy equivalents per car per year) or about 5–8 million homes. Globally, the production figure is around 40.5 million tonnes (equal to about 44.4 million short tons or about 475 billion cubic feet) and is expected to grow 3.5% annually through 2013 [2]. Major current uses of the commercially produced hydrogen are for oil refining, where hydrogen is used for hydrotreating of crude oil as part of the refining process to improve the hydrogen-to-carbon ratio of the fuel; food production (e.g., hydrogenation); treating metals; and producing ammonia for fertilizer and other industrial uses.

A range of potential hydrogen production methods and pathways is presented below in Fig. 1. These include reformation, gasification, electrolysis, and other advanced processes. In addition to these more well-established pathways, there are various emerging production pathways also possible. These are mostly at the laboratory scale and present and include such concepts as direct "photoelectrochemical" water splitting; a wide array of biological processes involving biogas conversion, algae, and microbial electrolysis cells; and other novel methods such as high-temperature thermochemical systems.

This chapter reviews a variety of established and emerging hydrogen production methods and their current status. Key remaining issues and research and development (R&D) challenges are highlighted for many of the pathways.

Review of Hydrogen Production Methods

Hydrogen in molecular form can be produced from many different sources and in many different ways. Hydrogen is most typically produced today for industrial uses through the steam reformation of natural gas



Hydrogen Production Science and Technology. Figure 1
Example hydrogen production pathways (Source: [1])

(known as “steam methane reforming”) and from other hydrocarbons in oil refinery locations, but also is produced through electrolysis of water and as a by-product of some industrial processes such as chlor-alkali production. While hydrogen production methods often have several aspects and can be grouped in various ways, they can be organized as follows for purposes of this entry:

- Reformation of fossil hydrocarbons and bioderived liquids
- Gasification of coal and other hydrocarbons
- Electrolysis of water
- Biomass gasification and pyrolysis
- Biological production
- Nuclear-related pathways
- Photoelectrochemical methods
- High-temperature thermochemical methods
- Fuel-cell-based coproduction

These production pathways are briefly reviewed below, along with identification of key remaining challenges associated with the ones that are not yet commercially viable. A section that compares production cost estimates of hydrogen by several of the methods is then presented followed by the overall conclusions of the chapter.

Reformation of Fossil Hydrocarbons and Bioderived Liquids

The most widely used hydrocarbon reformation process is steam methane reforming (SMR). SMR is the process by which natural gas, other hydrocarbon, or other methane-rich stream such as biogas or landfill gas is reacted with steam in the presence of a catalyst to produce hydrogen and carbon dioxide. In the SMR process, carbon monoxide is first produced with hydrogen (this blend is known as “syngas”), and then, a series of water–gas shift reactions converts the carbon monoxide to carbon dioxide and additional hydrogen through the injection of additional steam. The water–gas shift occurs first in a high-temperature step at about 350°C and then in a lower-temperature step at about 190–210°C. Four hydrogen molecules are produced for every one carbon dioxide molecule, making SMR a relatively low-carbon process when, for example, the use of SMR-based hydrogen is used in a fuel cell vehicle compared with the use of gasoline in a combustion-engine vehicle on a “full fuel cycle” basis [3, 4].

The overall SMR reactions are as follows:

$$\mathrm{CH_4 + H_2O\ (steam) \rightarrow CO + 3H_2\ (syngas\ production)}$$

$$\mathrm{CO + H_2O\ (steam) \rightarrow CO_2 + H_2\ (water\ gas\ shift)}$$

$$\mathrm{Overall: CH_4 + 2H_2O \rightarrow CO_2 + 4H_2}$$

SMR produces a hydrogen-rich gas that is typically on the order of 70–75% hydrogen on a dry basis, along with smaller amounts of methane (2–6%), carbon monoxide (7–10%), and carbon dioxide (6–14%) [5]. The efficiency of the SMR process using natural gas as a feedstock is typically about 74% on an LHV basis (CK) [6].

Bioliquids such as ethanol, sugars, and bio-oils can also produce hydrogen through reformation processes as well as via gasification and pyrolysis. Anderson and Carole [7] provide a useful review of the key advantages and disadvantages of various feedstocks as well as potential feedstock prices, hydrogen yield rates, and commercial production time frames. They examine ethanol, glucose, glycerol, sorbitol, ethylene and propylene glycol, cellulose/hemicellulose, methanol, and bio-oils as potential hydrogen feedstocks that are viable for use in reformation processes. They find that ethanol, glucose, glycerol, and bio-oils are potentially viable in the 2012–2017 time frame, while sorbitol, ethylene and propylene glycol, cellulose/hemicellulose, and methanol are possible in the 2017+ longer-term time frame. For the medium-term feedstocks, they find that the highest yield rates are possible for ethanol (0.26 kg hydrogen/kg ethanol) and crude glycerol (0.24 kg hydrogen/kg crude glycerol), with lower rates possible for pure glycerol (0.15 kg hydrogen/kg pure glycerol), glucose (0.13 kg hydrogen/kg glucose), and bio-oils (0.13 kg hydrogen/kg bio-oil). The longer-term options all have similar yield rates, ranging from 0.22 kg/ hydrogen/kg of feedstock (for methanol and propylene glycol) to 0.13/ kg hydrogen/kg for sorbitol and cellulose/hemicellulose, while ethylene glycol has a more intermediate estimated hydrogen yield of 0.15 kg per kg of feedstock [7]. They do not present overall levelized costs of hydrogen from each of the feedstocks, but based on the feedstock prices and yield rates presented, it appears that glucose, crude glycerol, and cellulose/hemicellulose could be particularly attractive from a rough feedstock cost and yield rate perspective (ignoring differences in capital equipment and process energy required, etc.).

US DOE production (not delivered) target costs are for \$3.80/gal equivalent for hydrogen from bioderived liquids in 2012 and under \$3.00/gal by 2017. Key technical challenges for achieving these goals include advances in catalyst systems, purification technologies, and system integration. A key issue is catalyst deactivation due to coking, when products from side reactions become deposited on the catalyst. An additional important area of research is low-temperature (<500°C) reforming, including aqueous-phase reforming that can be applied to glucose, ethylene glycol, sorbitol, glycerol, and methanol. Advantages include reduced energy intensity, better characteristics for the water–gas shift reactions, minimization of decomposition reactions (that can lead to the coking issue described above), and compatibility with membrane separation/purification techniques [7].

Gasification and Pyrolysis of Coal and Other Hydrocarbons

In the partial oxidation (POX) process, also known more generally as "gasification," hydrogen can be produced from a range of hydrocarbon fuels, including coal, heavy residual oils, and other low-value refinery products. The hydrocarbon fuel is reacted with oxygen in a less than stoichiometric ratio, yielding a mixture of carbon monoxide and hydrogen at 1200–1350°C. Hydrogen can also be produced through pyrolysis-based hydrocarbon gasification processes in the absence of oxygen.

The chemistry of hydrogen production from gasification of coal and other heavy hydrocarbons is complex, involving molecular cracking, POX, steam gasification, water–gas shift, and methanation reactions. The first stages involve progressive devolitization of the feedstock with increasing temperature, yielding a mix of oils, phenols, tars, and light hydrocarbon gases. This is followed by water–gas shift reactions, where additional hydrogen along with carbon dioxide are produced from carbon monoxide and water, and methanation reactions where methane (which can be later reformed into additional hydrogen) instead of hydrogen is formed with any remaining carbon monoxide. Depending on temperature and pressure conditions, the "syngas" produced from hydrocarbon gasification can have varying amounts of carbon dioxide, methane, and water, along with other trace components [8]. Several metals and metal oxides have been found to catalyze the gasification reactions and can alter the kinetic constants. Iron oxides such as Fe_2O_3 and Fe_3O_4 affect the rates of steam gasification of coal and other hydrocarbons, and some gasification reactions are also catalyzed by nickel and calcium-based substances [8].

One recent study [9] examines the potential life cycle impacts of hydrogen produced through gasification and pyrolysis of coal, along with two other pathways – the thermochemical sulfur–iodine cycle and the thermochemical "Westinghouse" cycle that both make use of solar thermal energy (or other source of waste heat such as from nuclear power) and sulfuric acid to produce hydrogen. These other chemical cycles are discussed in more detail in the thermochemical and nuclear-power-assisted hydrogen production sections later in this entry. The study uses SimaPro™ life cycle analysis software for the analysis, which includes major indices for human health impacts, ecosystem impacts (including air, water, and soil impacts), and resource use.

The study finds that, of these four methods of hydrogen production, from a life cycle (system construction and operation) perspective, gasification of coal has the lowest human health impacts and almost as low ecosystem quality impacts as the Westinghouse cycle, while the Westinghouse cycle has the lowest resource impacts. The sulfur–iodine cycle has somewhat higher ecosystem impacts, comparable human health impacts to the Westinghouse cycle, and comparable resource impacts to coal gasification (i.e., somewhat higher than for the Westinghouse cycle). Pyrolysis of coal, meanwhile, has by far the highest impacts across all three categories, primarily because the amount of hydrogen produced through coal pyrolysis is about 3.5 times lower than for gasification per unit of coal used as feedstock [9].

An interesting aspect of gasification is that it could lend itself to carbon dioxide capture, owing to the large scale of many gasification systems and the production of carbon dioxide through the water–gas shift reactions. Particular attention has focused on calcium oxide (CaO) as an inexpensive carbon dioxide acceptor, yielding calcium carbonate that could then be sequestered [9]. This could prevent the product carbon dioxide from entering the atmosphere, if appropriate geological or oceanic reservoirs are available in the area for long-term storage.

Electrolysis of Water

Electrolysis is the process through which water molecules are split into hydrogen and oxygen molecules using electricity and an electrolyzer device. Electrolyzers are, effectively, fuel cell devices that operate in reverse. The overall electrolysis reaction is:

$$e^- + H_2O \rightarrow {}^1\!/\!{}_2O_2 + H_2 \text{ (``water splitting'')}$$

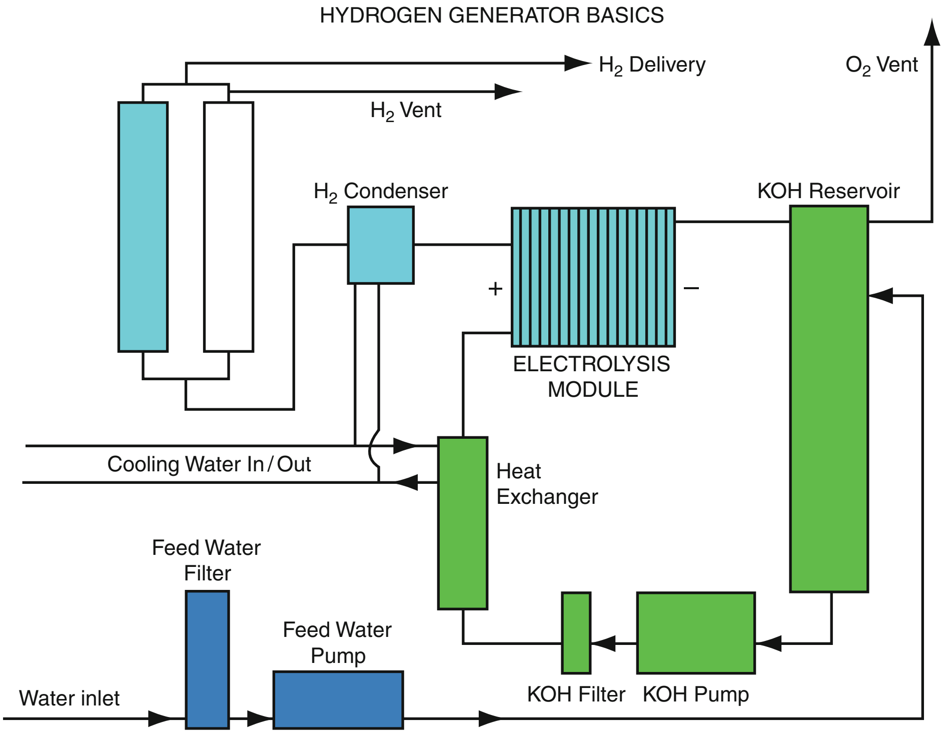
The two most common types of electrolyzers are alkaline (using a potassium hydroxide electrolyte) and PEM (using a solid polymer membrane electrolyte). A schematic of an alkaline electrolysis system is provided in Fig. 2. The electrolysis reaction produces pure oxygen as a by-product along with pure hydrogen. The oxygen can then be used for productive purposes such as enriching the oxygen content of greenhouses for food production.

Hydrogen can be produced via electrolysis of water from any electrical source, including utility grid power, solar photovoltaic (PV) power, wind power, hydropower, or nuclear power. The "Electrochemical Hydrogen Production" section of this encyclopedia reviews the basic science and materials issues with electrochemical hydrogen production in detail.

Electrolysis is currently done at a wide range of scales, from a few kilowatts to up to 2,000 kW per electrolyzer device. Additional research involves high-temperature electrolysis, where heat that is coproduced from another process, such as nuclear power as discussed below, is used to generate steam that can then be electrolyzed into hydrogen and oxygen with less electrical energy input than with lower-temperature electrolysis.

Biomass Gasification and Pyrolysis

Biomass conversion technologies can be divided into thermochemical and biochemical processes. Thermochemical processes tend to be less expensive because



Hydrogen Production Science and Technology. Figure 2
Hydrogen production by alkaline electrolysis (Courtesy of teledyne energy systems)

they can be operated at higher temperatures and therefore obtain higher reaction rates. They involve either gasification or pyrolysis (heating biomass in the absence of oxygen) to produce a hydrogen-rich stream of gas known as “syngas” (a blend of hydrogen and carbon monoxide). They can utilize a broad range of biomass types. In contrast, enzyme-based fermentative-type processes (discussed below) are at present mainly limited to wet, sugar-based feedstocks but could include cellulosic feedstocks in the future with continued improvements in process techniques and systems.

Biomass feedstocks can be gasified in a similar manner as fossil hydrocarbons as discussed above and can also be converted through pyrolysis. Gasification involves POX and/or steam reforming to produce a combination of gas and char products, where the char can then be reduced to ultimately produce a mix of hydrogen, carbon monoxide, carbon dioxide, and methane. Water–gas shift can then be used to produce additional hydrogen from carbon monoxide and water, and the methane can also be further reformed into additional hydrogen using SMR. Example chemical reactions include these, using glucose as an example hydrocarbon [10]:

$$C_6H_{12}O_6 + O_2 + H_2O \rightarrow CO + CO_2 + H_2 + \text{other species}$$

$$CO + H_2O \rightarrow CO_2 + H_2 + \text{small amount of heat(“water gas shift”)}$$

A key issue with biomass gasification is the formation of tar that occurs during the process, which is an unwanted product because it cannot be readily reformed into hydrogen. In order to minimize tar formation, careful design of the gasifier, incorporation of additives or catalysts, and careful control of operation variables can be effective strategies [8].

Biomass can also produce oxygenated oils through pyrolysis (without oxygen input), which can then be cracked and steam reformed to produce hydrogen and carbon dioxide syngas. An advantage of this method is that bio-oils are easier to transport than either biomass or hydrogen, and hydrogen can thus be produced in a two-step process at different locations with optimization of feedstock supply and hydrogen distribution [11].

Depending on application, there are potential additional cleanup steps (with associated costs) that may need to be added for hydrogen produced from biomass for the provision of high-purity hydrogen (e.g., for low-temperature fuel cell applications for transportation and stationary markets). These may be similar to those required for hydrogen production from other “fossil” hydrocarbons, but with considerable variability in both cases depending on the nature of the feedstock (i.e., natural gas or refinery coke), the amount of sulfur present, and other factors related to details of the production process.

Biological Production Methods

There are a wide range of additional biological production methods, beyond gasification or pyrolysis and reformation of biomass and reformation of biogases. These include:

- Photolytic hydrogen production from water
- Photosynthetic bacterial hydrogen production including dark fermentative methods
- Hydrogen production from algae
- Microbial electrolysis cells
- Other concepts such as combined biological systems

The primary of these additional biological production methods (the first four from the above list) are briefly discussed below.

Photolytic Hydrogen Production from Water

The photolytic hydrogen process produces hydrogen from water directly, without the use of a more conventional electrolyzer device. This is accomplished through the use of a photoelectrochemical (PEC) cell that employs a semiconductor material and involves movement of electrons through the material. This is similar to the operation of solar photovoltaic systems but where instead of the current being collected, it is immediately used at the site of the PEC to split water molecules in either an aqueous or flat-plate type of system.

In the photoelectrochemical water-splitting process, discussed in detail in the “► Photo-catalytic Hydrogen Production” entry in this encyclopedia [12], a photoanode is subjected to solar radiation, generating an electronic charge and an electron–hole pair (as in photovoltaics). Second, however, oxidation of water occurs at the holes of the anode, yielding molecular oxygen. Then, hydrogen ions and electrons are transported from the photoanode by way of an electrolyte (for the hydrogen ions) and an electrical connection (for the electrons), forming molecular hydrogen as the hydrogen ions are reduced with the electrons at the cathode. PEC cells can thus use sunlight as the only energy input to produce hydrogen [13].

Since only some of the solar spectrum involves photons with sufficient energy to generate these voltage levels, multilayer photovoltaic junction materials can be used to allow multiple low-energy photons to be grouped together to exceed the electrolysis threshold. For example, early research investigated $GaInP_2$/GaAs "p/n, p/n tandem cells" at the National Renewable Energy Agency, that incorporated a top $GaInP_2$ junction with a band gap of 1.83 eV, to absorb the visible part of the light spectrum, with a bottom GaAs junction with a band gap of 1.42 eV to absorb the near-infrared portion that gets transmitted through the top junction. This type of system was shown to have a hydrogen production efficiency of 12.4%, based on the estimated hydrogen yield and the incident light intensity of 1190 mW/cm^2, consistent with theoretical calculations showing a 24% maximum efficiency but more like 10–16% practical efficiencies for two-layer devices of this types [13].

The selection of semiconductor materials for the photoanode is critical for PEC cells, as discussed extensively in the ► Hydrogen via Direct Solar Production contribution to this encyclopedia [14]. Key issues identified in the chapter are for the semiconductor to have the proper band gap (the level of energy in electron volts needed, between about 1.6 and 2.0 eV for the primary cell) and the ability to avoid immediate recombination of the charge carriers.

Challenges to PEC-cell-based hydrogen production include materials issues including corrosion problems and uncertain performance over time, as the lifetimes of current cells being tested in laboratories are on the order of hundreds of hours where tens of thousands of hours of operation are needed for practical systems [15, 16]. Furthermore, there are challenges with the physical assembly of the supramolecular complexes needed for PEC cells, in order to make sure that the various sequential reactions needed are not restricted at any step by poor collection and distribution to the water substrate of the reducing equivalents. It is also important to note that understanding the factors that control multi-electron photochemistry is still at an early stage, meaning both that there are knowledge deficiencies as well as opportunities for future system improvements [14].

Practical PEC systems of several different designs can be contemplated, and four of these have undergone recent economic analysis for the US Department of Energy for an example 10 t/day hydrogen production plant. Of the four types studied – single-bed aqueous suspension, dual-bed aqueous suspension, fixed flat panel, and tracking concentrator – the first type was estimated to have the lowest costs of hydrogen produced ($1.63/kg) compared with $3.19/kg for the dual-bed suspension system, $4.05/kg for the tracking concentrator system, and $10.36/kg for the fixed flat panel system [17]. Given the early stage of development of PEC cells and systems for hydrogen production, however, these economic estimates should be treated as preliminary.

Photosynthetic Bacterial Hydrogen Production

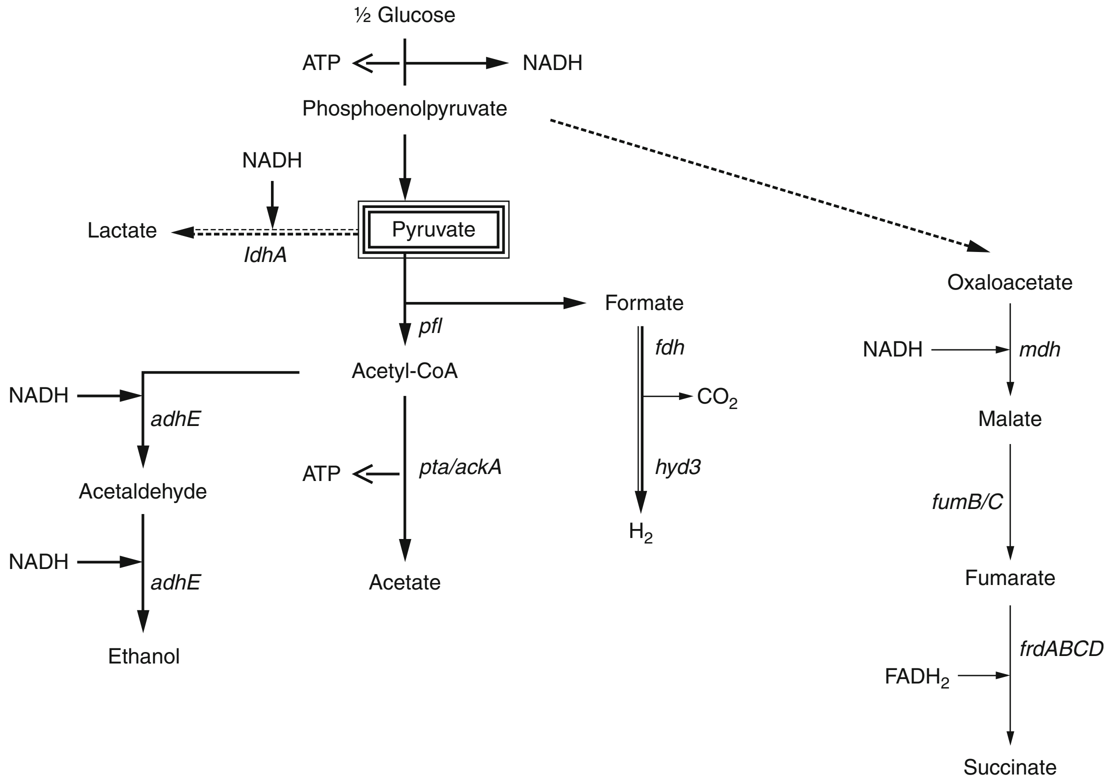
A number of fermentative processes for producing hydrogen using bacteria are being investigated. These can generally be categorized into "photobiological" methods where solar energy is captured to help drive the fermentation process, or those that use organic materials that were previously produced from photosynthetic process contain hydrogen that is associated with the fixation of carbon. Key issues with the direct photobiological methods are the relatively low conversion efficiencies of solar systems and the fact that oxygen is produced along with hydrogen, creating safety risks as well as difficulties in regulating the oxygen level in the fermentative environment (an important factor to the efficacy of these systems) [18].

Another option that involves conversion of existing organic materials to hydrogen is known as anaerobic or "dark" fermentation. This process uses some type of hydrogenase as an enzyme (the two main types being [NiFe] and [FeFe] hydrogenases), uses pyruvate as a key intermediate in the fermentation pathway, and works best with carbohydrate-rich feedstocks. Fermentation of proteins does not yield much hydrogen because protein amino acids are mostly fermented in pairs through "Strickland" reactions where one amino acid is an electron receptor for oxidation of the second amino acid, without hydrogen production [18]. Fermentation of carbohydrate-rich materials such as glucose can yield a series of products following the pyruvate intermediary, including formate that yields hydrogen and Acetyl-CoA that yields acetate and acetaldehyde that can be further processed into ethanol with the addition of NADH. This is known as "enteric-type mixed-acid fermentation" as shown in Fig. 3 below.

As with other biological hydrogen production methods that tend to be relatively slow in terms of yields, low yield rates for practical systems are a constraint for photosynthetic/bacterial methods as well. Based on known metabolic pathways, thermodynamic and metabolic limitations restrict hydrogen production to a maximum of 4 moles of hydrogen per mole of glucose [18]. This is because acetate is necessarily produced as well as hydrogen and carbon dioxide, effectively "locking up" half of the available hydrogen. Strategies to increase yields include optimizing parameters for the fermentation bioprocesses, attempting to limit competing reactions in order to channel more reductant for hydrogen production, decreasing hydrogen partial pressures in the reactor vessels (but then leading to more dilute streams of product gas), using thermophilic strains of bacteria that increase output at higher temperatures (but then requiring additional reactor vessel heating), and identifying and employing more active hydrogenases as enzymes [18].

Hallenbeck [18] reviews these strategies in some detail and concludes that it will be difficult to exceed the "Thauer limit" of 4 moles of hydrogen produced per mole of glucose, even though it is theoretically possible to produce 12 moles of hydrogen per mole of glucose by the reaction:

$$C_6H_{12}O_6 + 6H_2O \rightarrow 12H_2 + 6CO_2$$



Hydrogen Production Science and Technology. Figure 3
Enteric-type mixed-acid fermentation for hydrogen production (Source: [18])

The existing conversion efficiency of only 33% (maximum) of this theoretical limit is too low for practical systems unless additional steps can be employed to convert more of the organic feedstock into useful products. These have been proposed in the forms of a second stage fermentation for methane production, a second stage photofermentation for additional hydrogen production, and (more radically) metabolic engineering to develop novel hydrogen-producing pathways. Active research continues to address the current limitations, with the goal of developing more efficient and complete conversion levels in practical systems.

Hydrogen Production from Algae

The ability of various strains of algae to produce hydrogen has been studied for some time, with particular attention on unicellular green algae and potential combination of hydrogen production from green algae with additional hydrogen production from bacteria in a coproduction concept. Melis and Menecki [19] have proposed this type of system, where photosynthesis is employed in the first step to produce algae biomass and hydrogen gas using the [Fe]-hydrogenase enzyme. In the second stage, the green algal process is coupled with production of an anoxygenic bacteria (such as *Rhodospirillum rubrum*) to produce adenosine triphosphate (ATP), which is required for the evolution of hydrogen using the nitrogenase enzyme. Finally, during a third stage, fermentative bacteria such as *Clostridium pasteurianum* catabolize the algae/bacteria biomass and further enhance the hydrogen yield [19].

This type of three-organism system thus couples algal hydrogen production with a similar dark fermentative reaction step similar to the bacterial production process described above. Key areas of continuing research with algal-based hydrogen production systems include limiting oxygen evolution during the algal growth phase, exploring the role of sulfate availability to the chloroplasts to regulate the rate of photosynthesis, and examining various means for further increasing hydrogen yields to enable more practical hydrogen production systems [19].

Microbial Electrolysis Cells

In a related concept to photosynthetic biological production, microbial electrolysis cells (MECs) can be used to produce hydrogen by providing some of the voltage potential needed for electrolysis, thus reducing energy requirements. Experiments have determined that the bacteria can produce an anode working potential of around 0.3 V and that only an additional 0.11 V are needed to produce hydrogen in theory – but that in practice, more like an additional 0.25 V are needed due to overpotential at the cathode [20]. This means that it appears that approximately halving the energy needed for electrolysis is possible with MECs, but with still unexplored overall system efficiency. The complete system includes maintaining MEC operating conditions, delivering feedstocks to the production facility, replacing other expendable materials, and performing any gas separation and cleanup needed.

The MEC-based production concept requires two key steps. First, one group of bacteria turns unused sugar and unwanted vinegar from improper fermentation into electricity. Only a small amount of electricity is produced, however, and not enough to reach the 1.2 V necessary to split water in a typical electrolysis reaction. Therefore, some additional electricity from the power grid is also used. Second, another group of bacteria uses the electricity to split water molecules into oxygen and hydrogen in what is known as "microbial electrolysis" [20].

One of the biggest problems that practical systems have to overcome is the bacteria variability of the runoff water, making production rates difficult to predict because the bacteria have to build to a certain level of concentration to be effective. Another issue is that much of the hydrogen can be consumed by "methanogenic" microbes before leaving the solution, leading to much greater production of methane than hydrogen. While this methane could then be "reformed" into hydrogen by SMR, direct production would be far preferable from an overall energy-use standpoint. Research is underway in the laboratory to improve this hydrogen-to-methane production ratio [20].

Nuclear-Related Pathways

Many nuclear-energy-based hydrogen production schemes are possible. These include nuclear thermal conversion of water using various chemical processes such as the sodium-iodine cycle, electrolysis of water using nuclear power, and high-temperature electrolysis that additionally would use nuclear system waste heat to lower the electricity required for electrolysis. The basic concept is that heat from nuclear plants can be used to assist the production of hydrogen in various ways, with over 100 identified cycles that could benefit from association with nuclear power [21].

A recent US national laboratory study examined hydrogen production costs from nuclear cycles based on the sulfur–iodine (S–I) thermochemical water-splitting cycle, a hybrid sulfur (HyS) thermochemical water-splitting cycle (a modified Westinghouse-type cycle), and a high-temperature steam electrolysis (HTSE) cycle. The study focused on various sensitivities associated with cost estimates for these systems and generally found somewhat higher expected costs than have been reported previously (see Table 1 for details).

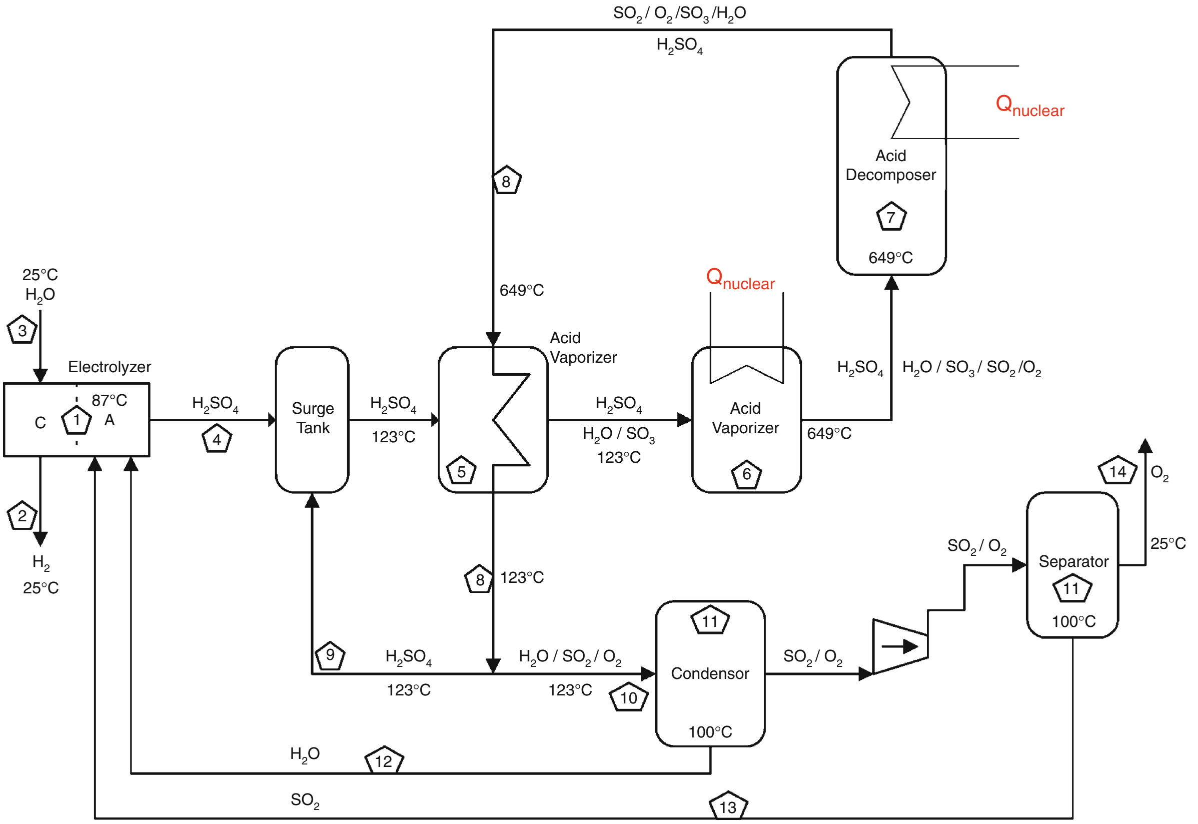
The S–I cycle has been studied since the 1970s and involves a three-step chemical reaction process. The first produces HI and sulfuric acid (H_2SO_4) from I_2, SO_2, and water at about 120°C; the second produces SO_2, water, and O_2 from the H_2SO_4 at about 830°C (assisted with heat from nuclear power); and the third produces I_2 and H_2 from the initial HI at about 450°C [22]. The sulfur and iodine compounds are recovered and reused. Based on a reactive distillation design, hydrogen production costs of about \$7.27/kg are believed possible with this design with a system producing 380 t/day of hydrogen, assuming some future system cost reductions [23].

The HyS design differs from S to I in that it uses thermal input energy for oxygen generation and a separate low-temperature electrolysis step for hydrogen generation, but also involves sulfuric acid decomposition. The HyS cycle was first explored by Westinghouse in the 1980s and is sometimes called the “Westinghouse hybrid cycle.” The key steps in the Westinghouse HyS cycle are [24]

Hydrogen Production Science and Technology. Table 1 Estimated hydrogen production costs

Technology and fuel	Production capacity (1,000 kg/day)	Hydrogen production cost (\$/kg)				Source
		Capital (\$)	Feedstock (\$)	O&M (\$)	Total (\$)	
Central SMR of natural gas	380	0.16	1.15	0.14	1.47	[1]
Distributed SMR of natural gas	1.5	0.40	1.72	0.51	2.63	[1]
Central coal gasification with CCS	308	0.83	0.56	0.43	1.82	[1]
Central coal gasification without CCS	284	0.57	0.56	0.09	1.21	[1]
Biomass gasification	155	0.37	0.52	0.55	1.44	[1]
Distributed electrolysis	1.5	0.96	5.06	0.73	6.75	[1]
Central wind electrolysis	124.5	1.46	1.69	0.65	3.82	[1]
Distributed wind electrolysis	0.5	3.00	3.51	0.74	7.26	[1]
Central nuclear thermochemical	1,200	0.76	0.20	0.43	1.39	[1]
Nuclear S–I	343.5	2.61	0.35	5.31	7.27	[23]
Nuclear HyS	343.5	1.48	0.22	2.25	4.95	[23]
Nuclear HTSE	343.5	1.05	0.03	3.15	4.23	[23]
Single-bed aqueous PEC	1	–	–	–	1.63	[17]
Dual-bed aqueous PEC	1	–	–	–	3.19	[17]
Fixed panel PEC	1	–	–	–	4.05	[17]
Tracking concentrator PEC	1	–	–	–	10.36	[17]

Notes: *CCS* carbon capture and sequestration, *HTSE* high-temperature steam electrolysis, *HyS* hybrid sulfur cycle, *O&M* operations and maintenance, *PEC* photoelectrochemical cell, *S–I* sulfur–iodine, *SMR* steam methane reforming



Hydrogen Production Science and Technology. Figure 4
Westinghouse hybrid sulfur cycle schematic (Source: [24])

$$H_2SO_4 \rightarrow H_2O + SO_3 \rightarrow SO_2 + 1/2O_2 + H_2O \text{ (thermal decomposition, } 871^\circ\text{C)}$$

$$2H_2O + SO_2 \rightarrow H_2SO_4 + H_2 \text{ (electrolysis, } 87^\circ\text{C)}$$

Hydrogen production costs of about $4.95/kg are believed possible with this type of hybrid system design with a system producing 345 t/day of hydrogen [23]. A schematic of the HyS cycle is presented in Fig. 4:

The third design explored in the recent study, HTSE, involves electrolysis of water at about 800°C, assisted with heat from nuclear power systems. Key elements of the system include sweep gas coupling heat exchangers, process coupling heat exchangers, solid oxide electrolyzer systems, and feed and product purification systems. Hydrogen production costs of about $4.23/kg are believed possible with HTSE with a system producing 345 t/day of hydrogen [23].

Thermochemical Production Methods

In this set of pathways, high-temperature (500–2,000°C) heat is used to drive chemical reactions to produce hydrogen, including via direct thermolysis of water. This can be done in conjunction with solar thermal power, nuclear power, or other heat sources. Of the over 200 pathways that have been identified, several are being pursued by US DOE for further research [25].

Thermochemical production typically involves the use of metal oxides in a two-step process where a metal oxide is first decomposed into the metal and oxygen in an endothermic step, and then, the metal is then combined with water to re-form the metal oxide and hydrogen in a second exothermic step [26]. In one example known as the zinc oxide cycle, zinc oxide powder passes through a reactor heated by a solar concentrator operating at about 2,300 K. At this temperature, the zinc oxide dissociates to zinc and oxygen gases. The zinc cools, separates, and reacts with water to form hydrogen gas and solid zinc oxide:

$$2ZnO + heat \rightarrow 2Zn + O_2$$

$$2Zn + 2H_2O \rightarrow 2ZnO + 2H_2$$

The net result is hydrogen and oxygen, produced from water, where the hydrogen can be separated and purified for subsequent use. The zinc oxide can be recycled and reused to create more hydrogen through subsequent phases of the process. Key limitations at present include the rapid reoxidation of the zinc vapor and challenges with keeping the receiver windows clean when the system is coupled with solar thermal systems for delivery of the heat required [26].

An additional type of thermochemical system has been proposed that would also involve a photon component, essentially coupling thermochemical and photoelectrochemical hydrogen production pathways. One concept would use a carbon dioxide/carbon monoxide cycle, where carbon dioxide molecules would absorb near-ultraviolet photons and high temperatures and dissociate into carbon monoxide, helping to enable the production of hydrogen. Research has showed that this can be done at temperatures of around 2,000–2,500 K, which would be compatible with solar thermal concentrator systems [26].

Other concepts involve modifications to the Westinghouse cycle discussed above, including a proposed sulfur–ammonia cycle that makes use of solar power in two ways –to drive both a solar photocatalytic reactor and a sulfuric acid decomposition reactor. The process would use ammonia, sulfuric acid, and aqueous ammonium sulfate – (NH4)2SO4 – at various stages, with the principal advantages using solar photons to directly produce the chemical energy of hydrogen along with using solar heat in the thermochemical step, thus eliminating the need for electrical energy input [26].

Fuel Cell Coproduction

High-temperature fuel cells based on molten carbonate (MCFC) or solid oxide (SOFC) technology operate at sufficiently high temperatures to run directly on methane. This is sometimes called "internal reforming." Thus, MCFC and SOFC systems do not need a pure or relatively pure hydrogen stream as do proton exchange membrane (PEM) and phosphoric acid (PAFC) systems, but can run directly on natural gas or biogas or landfill gas. Furthermore, such systems can be designed to produce additional purified hydrogen as a by-product (e.g., for use as a vehicle fuel), by feeding additional fuel and then purifying the hydrogen-rich "anode tail gas" from the fuel cell into purified hydrogen. This concept is discussed in some detail in the case studies section below. A detailed review of this concept for hydrogen production can be found in the "▶ Hydrogen Production from High-Temperature Fuel Cells" entry in this encyclopedia [27].

FuelCell Energy (FCE) of Danbury, Connecticut, has teamed up with Air Products and Chemicals, Inc., (APCI) to demonstrate renewable hydrogen production based on the FCE molten-carbonate fuel cell technology and a novel hydrogen gas cleanup system. This system, called the "DFC-H2®" system, has undergone initial testing at FCE's research laboratories and is now being readied for deployment at a landfill site at the Orange County Sanitation District in southern California [28]. Figure 5 below shows the system pictured during initial testing.

The basic concept behind the energy station is that hydrogen and electricity are coproduced, where the electricity is produced using methane (natural gas or biogas) as a feedstock in the high-temperature fuel cell, but additional hydrogen is produced within the fuel cell stack leading to a hydrogen-rich stream of gas leaving the fuel cell unit. This hydrogen-rich "anode tail gas" can then be purified for other uses, such as fuel cell vehicles or other types of fuel cells (e.g., PEM and PAFC) that require pure hydrogen. The following schematic presents the concept and how there are also opportunities for waste heat recovery to help boost overall efficiency.

FCE and APCI have recently released a technical paper about the system that suggests that 125 kg/day of hydrogen can be produced along with an electrical output of 250 kW, based on over 8,500 h of system testing and a "pressure swing adsorption" (PSA) process for hydrogen separation. The purity of the product hydrogen was measured at 99.99%, while also meeting a target of 0.2 ppm of carbon monoxide.

The company believes that a novel electrochemical hydrogen separation unit, which they are also testing, could offer up to a 50% reduction in operating cost compared to the more conventional PSA unit while

Hydrogen Production Science and Technology. Figure 5
"DFC-H2®" system being tested prior to field deployment (Source: [28])

offering an overall electrical power plus hydrogen production efficiency increase to 68% from 66% (with no waste heat recovery) on a lower-heating-value basis [28].

Hydrogen Production Costs

Table 1 presents a set of hydrogen-production-only costs (i.e., not including delivery if centralized production) by some of the production methods discussed in this entry that are either used at present or that are possible in the future. These estimates are all as reported by the US Energy Information Administration or based on analysis for the National Academy of Science or the US Department of Energy. These and other estimates were used to report the production cost ranges by production methods that were discussed in the preceding sections of this entry. Note that some of the cleanest methods of producing hydrogen are currently the most expensive (e.g., based on electrolysis from wind or other clean electricity sources), but biomass gasification offers a renewable hydrogen pathway with costs that can potentially be competitive with fossil sources.

In comparison with these projections, in the current hydrogen market, deliveries of industrial-grade high-purity (99.95%+) hydrogen at medium to large volumes of hundreds to a few thousands of kilograms are typically priced at around $4–5 per 100 standard cubic feet (SCF) (or $17–21/kg), plus freight, rental, and hazmat charges. For smaller volumes of between 4 and 15 kg, or about 1,600–6,100 SCF, gas costs are estimated to vary widely from $5 to $20 per 100 SCF ($21–83/kg) depending on service model, delivery frequency, and distance from the gas supplier fill plant [29].

Conclusion

In conclusion, hydrogen is a promising energy carrier and fuel for stationary and transportation uses, but the potential expanded use of hydrogen involves many technical and infrastructure-related challenges. Approximately 10–11 million metric tonnes of hydrogen are used each year in the USA and about 40 million tonnes globally, but mostly in internal industrial settings. Dispensing hydrogen to wider consumer and fleet-use markets, such as for private and fleet vehicles and for stationary power uses, involves considerable challenges due to hydrogen's low energy density (by volume) and other unusual characteristics (i.e., small molecule size, relative ease of ignition in mixes with air, and need for high gas purity when used in certain applications such as in PEM fuel cells).

There are many available and emerging hydrogen production pathways based on a myriad of thermochemical, photochemical, and electrochemical mechanisms and reactions. Perhaps the most compelling aspect of hydrogen as a potential fuel is the incredible diversity in the means and methods by which it can be produced – literally by everything from basic algae to high-temperature fuel cells using ceramic materials and customized catalysts. Because hydrogen can be produced from such a wide range of feedstocks – from fossil sources such as natural gas or coal to a variety of renewable sources including those based on solar power and wind, geothermal, and biomass energy.

Options for hydrogen production include various pathways using fossil hydrocarbons or biogas using gasification or pyrolysis processes; steam methane reforming; electrolysis of water; nuclear-power-assisted cycles including the sulfur–iodine cycle; fermentative, algae-based, and other biological pathways; and more recently developed electrochemical and thermochemical processes including the use of PEC cells, microbial electrolysis cells, and tailored molecules that can facilitate the splitting of water molecules into hydrogen and oxygen with lower energy requirements than conventional electrolysis.

Of these, several pathways have been studied with regard to potential production costs at commercial-scale operations. While cost estimates are somewhat preliminary for some of the more novel production methods, particularly attractive hydrogen production costs appear possible for centralized SMR of natural gas, gasification of coal and biomass, nuclear thermochemical systems, and single-bed aqueous photoelectrochemical methods. Also potentially attractive are smaller-scale decentralized production systems based on SMR or electrolysis, as these can avoid the expense and difficulty of transporting bulky hydrogen gas or cryogenic liquid by generating the hydrogen close to the point of end use.

Bibliography

1. US Energy Information Administration (EIA) (2008) The Impact of Increased Use of Hydrogen on Petroleum Consumption and Carbon Dioxide Emissions, Report #: SR-OIAF-CNEAF/2008-04. www.eia.doe.gov/oiaf/servicerpt/hydro/appendixc.html, August
2. Freedonia Group (2010) World hydrogen demand and 2013 forecast. February
3. Lipman TE, Delucchi MA (2010) Expected greenhouse gas emission reductions by battery, fuel cell, and plug-in hybrid electric vehicles. In: Pistoia G (ed) Battery, hybrid, and fuel cell vehicles. Elsevier, Amsterdam. ISBN: 978-0-444-53565-8
4. Wang M, Wu Y, Elgowainy A (2007) GREET1.7 fuel-cycle model for transportation fuels and vehicle technologies. Argonne National Laboratory, Argonne
5. Hirschenhofer JH, Stauffer BD, Engleman RR, Klett MG (2000) *Fuel Cell Handbook*, Fourth Edition. Department of Energy, Federal Energy Technology Center, B/T Books, Orinda
6. US Department of Energy (2010) The Department of energy hydrogen and fuel cells program plan an integrated strategic plan for the research, development, and demonstration of hydrogen and fuel cell technologies – DRAFT
7. Anderson A, Carole T (2006) Bio-derived liquids to hydrogen distributed reforming working group background paper. U.S. Department of Energy, HFCIT Program, September
8. Navarro RM, Pena MA, Fierro JLG (2007) Hydrogen production reactions from carbon feedstocks: fossil fuels and biomass. Chem Rev 107:3952–3991
9. Smitkova M, Janicek R, Riccardi J (2009) Life cycle analysis of processes for hydrogen production. Int J Hydrog Energy 34(3):1370–1376
10. U.S. Department of Energy (2011) Hydrogen production: biomass gasification. Web document, http://www1.eere.energy.gov/hydrogenandfuelcells/production/biomass_gasification.html?m=1&
11. Czernik S, French R, Penev M (2010) Distributed bio-oil refining. DOE hydrogen program FY 2010 annual progress report, pp 38–41
12. Zhu J (2011) Photocatalytic hydrogen production, (this volume)
13. Khaselev O, Turner JA (1998) A monolithic photovoltaic-photoelectrochemical device for hydrogen production via water splitting. Science 280:425–427
14. Arachchige SM, Brewer KJ (2011) Hydrogen via direct solar production. (this volume)
15. Joshi AS, Dincer I, Reddy BV (2010) Exergetic assessment of solar hydrogen production methods. Int J Hydrog Energy 35:4901–4908
16. Brewer KJ, White T, Quinn K, Wang J, Arachchige S, Knoll J (2011) Photoinitiated electron collection in mixed-metal supramolecular complexes: development of photocatalysts for hydrogen production. DOE hydrogen and fuel cells program 2011 annual merit review meeting
17. James BD, Baum GN, Perez J, Baum KN (2009) Technoeconomic analysis of photoelectrochemical (PEC) hydrogen production. Prepared by Directed Technologies Inc. for the U.S. Department of Energy under DOE Contract Number: GS-10 F-009J, December
18. Hallenbeck PC (2009) Fermentative hydrogen production: principles, progress, and prognosis. Int J Hydrog Energy 34:7379–7389
19. Melis A, Melnecki MR (2006) Integrated biological hydrogen production. Int J Hydrog Energy 31:1563–1573
20. Logan BE, Regan JM (2006) Microbial fuel cells – challenges and applications. Environ Sci Technol 1:5173–5180
21. Brown LC, Besenbruch GE, Funk JE, Marshall AC, Pickard PS, Showalter SK (2002) High Efficiency Generation Of Hydrogen Fuels Using Nuclear Energy. Presentation at U.S., Department of Energy Hydrogen Fuel Cells and Hydrogen Review, Nuclear Energy Research Initiative (NERI)
22. Mathias PM, Brown LC (2003) Thermodynamics of the sulfur-iodine cycle for thermochemical hydrogen production. Presented at the 68th annual meeting of the society of chemical engineers. Japan, 23 March 2003

23. Allen D (2009) FY09 projected hydrogen cost estimates for nuclear hydrogen initiative baseline processes. SAND2009-6630P Sandia National Laboratory, Albuquerque
24. Jeong YH, Kazimi MS, Hohnholt KJ, Yildiz B (2005) Optimization of the hybrid sulfur cycle for hydrogen generation. MIT–NES–TR–004, May
25. US Department of Energy (2011) High-temperature water splitting. http://www1.eere.energy.gov/hydrogenandfuelcells/production/water_splitting.html
26. T-Raissi A, Muradov N, Huang C, Adebiyi O (2007) Hydrogen from solar via light-assisted high-temperature water splitting cycles. Trans ASME 129:184–189
27. Brouwer J, Margalef P (2011) Hydrogen production from high temperature fuel cells (this volume)
28. Patel P, Lipp L, Jahnke F, Heydorn E, Abdallah T, Holcomb F (2010) Co-production of renewable hydrogen and electricity: technology development and demonstration. ECS Trans 26(1):493–504
29. Cohen M, Snow GC (2008) Hydrogen delivery and storage options for backup power and off-grid primary power fuel cell systems. Published in IEEE Intelec 2008 proceedings

Hydrogen via Direct Solar Production

Shamindri M. Arachchige, Karen J. Brewer
Department of Chemistry, Virginia Polytechnic
Institute and State University, Blacksburg, VA, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Solar Hydrogen Photocatalysis
Conclusions and Future Directions
Acknowledgments
Bibliography

Glossary

bpy A bidentate terminal ligand, 2,2′-bipyridine.

dpp A bidentate bridging ligand, 2,3-bis(2-pyridyl) pyrazine.

dpq A bidentate bridging ligand, 2,3-bis(2-pyridyl) quinoxaline.

dpb A bidentate bridging ligand, 2,3-bis(2-pyridyl) benzoquinoxaline.

EA A subunit that functions to accept an electron in a redox reaction, electron acceptor.

ED A subunit that functions to donate an electron in a redox reaction, electron donor.

TL A ligand that is located at the end of a supramolecule and bound to only one metal center, terminal ligand.

BL A ligand that is located between two metals in a supramolecule, bridging ligand.

CAT The site of a reaction that occurs via catalysis, facilitated by the catalyst without the destruction of this subunit occurring multiple times in a cycle.

I_a Intensity of incident light.

et Excited state electron transfer, oxidation or reduction reaction involving the loss or gain of an electron that occurs in an electronic excited state.

GS Electronic ground state of a molecule.

HOMO Highest occupied molecular orbital for a molecule.

ic Internal conversion, a non-radiative process that allows conversion between two electronic states of the same spin multiplicity.

isc Intersystem crossing, a non-radiative process that allows conversion between two electronic states of different spin multiplicity.

k_x Rate constant of process “x.”

LA Ground electronic state of a light absorber.

*LA Excited electronic state of a light absorber.

LUMO Lowest unoccupied molecular orbital for a molecule.

Me$_2$bpy A bidentate terminal ligand, 4,4′-dimethy-2,2′-bipyridine.

Me$_2$phen A bidentate terminal ligand, 4,7-dimethylphenanthroline.

L-pyr A monodentate terminal ligand, [(4-pyridine) oxazolo(4,5-f)phenanthroline].

dmgH Dimethylglyoximate.

dmgBF$_2$ (Difluoroboryl) dimethylglyoximate.

dpgBF$_2$ (Difluoroboryl)diphenylglyoximate.

ppy A monodentate terminal ligand, 2-phenylpyridine.

py A N containing aromatic ring, pyridyl.

MLCT Metal-to-ligand charge-transfer, an electronic excited state.

MMCT Metal-to-metal charge-transfer, an electronic excited state.

nr Non-radiative decay or movement between electronic states via processes that do not involve the absorption or emission of light.

phen A bidentate terminal ligand, 1,10-phenanthroline.

q Bimolecular deactivation, a reaction of an electronic excited state of a molecule which results in relaxation to the ground state via interaction with another molecule without transfer of electrons or generation of an excited state.

Q Quencher, a molecule that is engaged in a reaction with an electronic excited state of another molecule leading to generation of the ground state of the excited molecule.

rxn Photochemical reaction from the electronic excited state of a molecule.

tpy A tridentate terminal ligand, 2,2′:6′,2″-terpyridine.

Φ^{em} Quantum yield of emission providing the number of photons emitted by an electronic excited state of a molecule divided by the number of photons absorbed.

Φ Quantum yield or a process provides the number of events of interest that occur from an electronic excited state as related to the number of excited molecules.

$\lambda_{max}{}^{abs}$ Absorption maximum.

$\lambda_{max}{}^{em}$ Emission maximum.

Definition of the Subject

Developing renewable sources of energy to replace reliance on fossil fuels remains a major challenge. Sunlight is recognized as the largest of all carbon-neutral energy sources with 4.3×10^{20} J reaching the earth's surface in 1 h, larger than the current energy consumption by humans in a year [1, 2]. Plants harness solar energy through photosynthesis, which is used to power the biological world and ultimately stored in carboniferous fossil fuels. The rate of production of fossil fuels is exceeded by current consumption leading to interest in artificial means to harvest solar energy. Growing energy demands due to rising living standards and increased population have stimulated a global initiative to exploring alternative means to harness solar energy and alternative energy sources. Current methods to harness solar energy for practical purposes do not compete with fossil fuel use. Artificial photosynthesis, use of synthetic systems that mimic photosynthetic energy conversion, provides an attractive avenue for solar fuel production [1–4]. The "hydrogen economy," transformation from an oil-based economy to a hydrogen-based one, has been proposed as an attractive energy solution for the future [1–6]. Hydrogen provides a high energy content per gram (120 kJ/g), can be used in fuel cells, and has low environmental impact. Hydrogen is found in various feedstocks on earth, combined with oxygen in water and carbon in fossil fuels. Steam-methane reforming currently provides about 95% of the hydrogen used in the United States [7]. Visible light-induced solar hydrogen production from water would be the ultimate clean, renewable energy solution. Molecule-based systems for artificial photosynthesis offers unique insight into electron transfer processes amenable to study by conventional experimental techniques. A molecular-based artificial photosynthetic fuel production system requires an antenna to capture light, generation of a potential, and a catalyst for reduction of the substrate, water. Water reduction to produce hydrogen fuel is complex involving charge separation, multielectron reactions, and chemical bond reorganization. Functioning molecular photocatalysts for water reduction have emerged recently, but practical systems do not yet exist. Engineering systems that can efficiently produce hydrogen from water is a topic of extensive research and developmental efforts with significant progress made in identifying the molecular components necessary for this complicated chemistry.

Introduction

This contribution highlights recent progress in the use of complex molecular systems that reduce water or protons for solar hydrogen production. Supramolecular complexes in this forum are large molecular assemblies composed of multiple components whose individual properties provide a unique function to the supramolecule [8]. The focus will be on supramolecular complexes that couple a light-absorbing unit (LA) to a reactive metal center that is the catalyst (CAT). Systems are designed so that the orbital energetics of the molecular components allow efficient intramolecular charge-transfer to the CAT upon light activation. Although the properties of the components are typically perturbed in the supramolecular assemblies, the individual components bring to the molecular device a unique function retained by each subunit in the assembly. A sacrificial electron donor (ED) is typically used to reductively quench the oxidized LA. The foundation for supramolecular photochemistry was established by the pioneering work using separate LAs and hydrogen-generating CATs in solar energy conversion schemes [9–11]. Recent progress using homogeneous multicomponent systems in solar hydrogen production will also be presented. The multicomponent systems consist of a LA, electron relay (ER), and a CAT that are not covalently linked. Illustrative examples of different molecular architectures are provided and the photochemistry discussed. Factors impacting system efficiency, including light-absorbing ability at desired wavelengths, the quantum yield of product formation (Φ), and the turnover number (TON) for system components, will be highlighted.

Photochemical Water Splitting

Solar water splitting uses energy from the sun to split water into hydrogen and oxygen. The splitting of water is an energetically uphill process requiring 1.23 eV. The overall reactions for water splitting are multielectron processes (Eqs. 1, 2, and 3):

$$2\,H_2O(l) \rightarrow 2\,H_2(g) + O_2(g) \quad E = -1.23\ \text{V vs NHE} \tag{1}$$

$$4\,H_2O(l) + 4e^- \rightarrow 2\,H_2(g) + 4OH^-(aq) \quad E = -0.83\ \text{V vs NHE} \tag{2}$$

$$4\,OH^-(aq) \rightarrow 2\,H_2O(l) + O_2(g) + 4e^- \quad E = -0.40\ \text{V vs NHE} \tag{3}$$

The redox potentials are pH dependent with proton reduction at pH = 1 occurring at 0.00 V vs. NHE. Multielectron chemistry is key as multielectron water is energetically much more favorable (1.23 V vs. NHE) [1] than water splitting using single electron pathways (5 V vs. NHE) [11]. Most of the solar spectrum has energies >1.23 eV and can be used for the multielectron splitting of water. Many systems are reported to split water using ultraviolet radiation but the fraction of available solar power in this region is smaller, and focus has been to use more abundant, lower energy light to power water photolysis [3, 4]. Due to the complicated nature of designing complete water splitting systems, researchers typically focus on understanding either water oxidation or water reduction, by replacing the other half with a sacrificial electron acceptor or a sacrificial ED.

Components for Supramolecular Solar Hydrogen Photocatalysis

The individual components within the supramolecule provide the overall photocatalytic activity to the supramolecule, thus careful consideration must be given to selecting structural components. The perturbation of each component upon assembly in a supramolecule must be predictable and controllable, with the overall function being maintained.

Light Absorption LA is used in this forum to absorb light and transfer that energy or an electron to other parts of the supramolecule. Incorporation of a LA within the molecular architecture that absorbs in the visible region allows for the use of more of the solar spectrum in solar hydrogen production schemes. A LA absorbs a photon of light to form an electronically excited state, *LA, which has distinctive properties compared to the ground state system, Eq. 4:

$$LA + h\nu \xrightarrow{I_a} {}^*LA \tag{4}$$

Once excited, relaxation of the *LA to the ground state can occur in a non-radiative fashion, through heat loss or in a radiative fashion by emitting light, Eqs. 5 and 6:

$${}^*LA \xrightarrow{k_{nr}} LA + \text{heat} \tag{5}$$

$${}^*LA \xrightarrow{k_r} LA + h\nu' \tag{6}$$

Excited states can be quenched by energy or electron transfer, see Eqs. 7, 8, and 9:

$${}^*LA + Q \xrightarrow{k_{en}} LA + {}^*Q \tag{7}$$

$${}^*LA + EA \xrightarrow{k_{et}} LA^+ + EA^- \tag{8}$$

$${}^*LA + ED \xrightarrow{k'_{et}} LA^- + ED^+ \tag{9}$$

The thermodynamic driving force (E_{redox}) of excited state electron transfer depends on the excited state reduction potential of the LA ($E(^{*}LA^{n+}/LA^{(n-1)+})$) and the ground state oxidation potential of the ED ($E(ED^{0/+})$) for excited state reduction, Eq. 10, or the excited state oxidation potential of the LA ($E(^{*}LA^{n+}/LA^{(n+1)+})$) and the ground state reduction potential of the EA ($E(EA^{0/-})$) for excited state oxidation, Eq. 11. The excited state redox potentials, $E(^{*}LA^{n+}/LA^{(n-1)+})$ and $E(^{*}LA^{n+}/LA^{(n+1)+})$, differ significantly from the ground state potentials as electronic excited states are both better oxidizing and reducing agents relative to their ground states. The excited state redox potentials can be calculated based on the ground state reduction or oxidation potential of the LA and the energy gap between the ground vibronic state of the electronic ground and excited states (E^{0-0}), Eqs. 12 and 13:

$$E_{redox}(red) = E\left(^{*}LA^{n+}/LA^{(n-1)+}\right) - E\left(ED^{0/+}\right) \tag{10}$$

$$E_{redox}(oxd) = E\left(^{*}LA^{n+}/LA^{(n+1)+}\right) - E\left(EA^{0/-}\right) \tag{11}$$

$$E\left(^{*}LA^{n+}/LA^{(n-1)+}\right) = E\left(LA^{n+}/LA^{(n-1)+}\right) + E^{0-0} \tag{12}$$

$$E\left(^{*}LA^{n+}/LA^{(n+1)+}\right) = E\left(LA^{n+}/LA^{(n+1)+}\right) - E^{0-0} \tag{13}$$

Common Light Absorbers Ruthenium and late transition metal complexes are widely applicable LAs with strong absorptivity and good spectral coverage as well as long-lived emissive metal-to-ligand charge-transfer (MLCT) excited states [9, 10, 12–15]. Some commonly used LAs used in solar energy conversion schemes are shown in Fig. 1. The excited state of $[Ru(bpy)_3]^{2+}$ is capable of excited state electron- [16–18] and energy-transfer reactions. $[Ru(bpy)_3]^{2+}$ absorbs light with high extinction coefficients in the ultraviolet and visible regions of the spectrum. The UV-region is dominated by intense absorptions that are intraligand (IL) $\pi \rightarrow \pi^{*}$ in nature, while the visible region is dominated by $Ru(d\pi) \rightarrow bpy(\pi^{*})$ MLCT transitions with $\lambda_{max}{}^{abs} = 450$ nm [12–14]. Following the absorption of a photon of light to generate the ^{1}MLCT excited state, rapid intersystem crossing occurs with unit efficiency to populate a relatively long-lived ^{3}MLCT excited state (excited state lifetime, $\tau = 850$ ns in room temperature (RT) acetonitrile solution, quantum yield, $\Phi = 0.07$ emitting at $\lambda_{max}{}^{em} = 620$ nm in RT) [13]. The excited state lifetime, τ, is the inverse of the sum of all the rate constants for deactivation of an electronic excited state in the absence of a quencher, Eq. 14:

$$\tau = \frac{1}{k_r + k_{nr}} \tag{14}$$

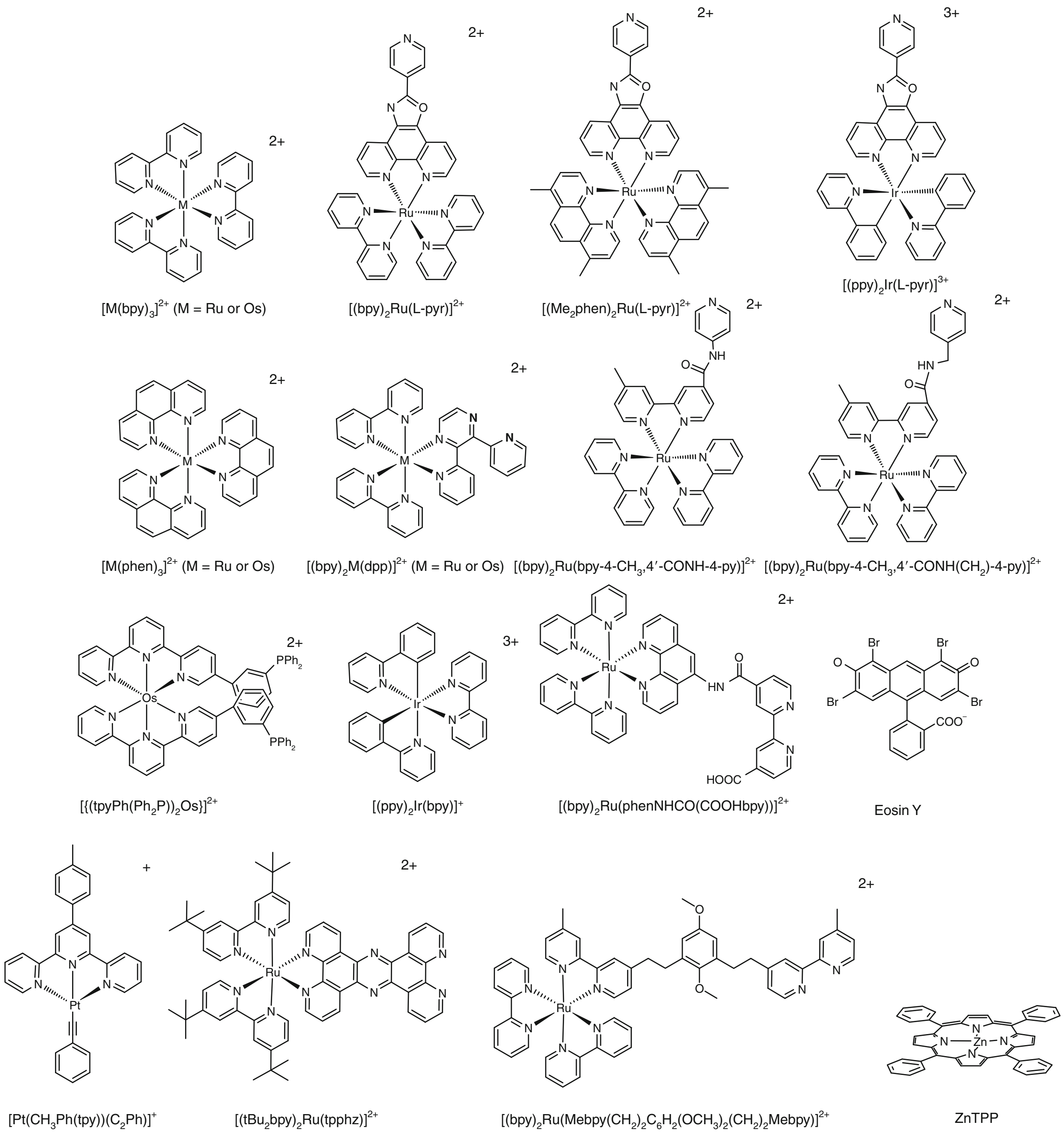
The quantum yield for emission, Φ^{em}, for an excited state can be calculated by considering the ratio between the rate constants for the process of interest and the sum of all the rate constants for the deactivation of a state, multiplied by the fraction of light that populates this state, Eq. 15:

$$\Phi^{em} = \Phi_{^{3}MLCT} \frac{k_r}{k_r + k_{nr}} \tag{15}$$

$\Phi_{^{3}MLCT}$ is the quantum efficiency for generation of the ^{3}MLCT state, which is one for $[Ru(bpy)_3]^{2+}$ and most Ru(II) polyazine LAs. The relationship between Φ of photochemical processes with concentration of a Q is described using Stern-Volmer kinetics, Eq. 16:

$$\Phi^{o}/\Phi = 1 + K_{sv}[Q] \tag{16}$$

Φ^{o} and Φ are quantum yields in the absence and presence of Q, respectively, and K_{sv} is the Stern-Volmer quenching constant. The excited state properties of Ru polyazine–type LAs can be modulated by modification of the ligands bound to the metal center. Complexes use similar properties with higher energy $Os(d\pi)$ orbitals and lower energy-excited MLCT states. Table 1 provides spectroscopic properties of Ru and Os LAs. Although the excited state of $[Ru(bpy)_3]^{2+}$ is thermodynamically capable of reducing water to produce hydrogen, this reaction does not occur. Recent studies have also used Ir- or Pt-based LA in photochemical hydrogen production schemes [19–21]. The use of metalloporphyrin types of complexes as LAs have been investigated as they mimic dyes found in nature in photosynthesis [22].



$[M(bpy)_3]^{2+}$ (M = Ru or Os)

$[(bpy)_2Ru(L\text{-}pyr)]^{2+}$

$[(Me_2phen)_2Ru(L\text{-}pyr)]^{2+}$

$[(ppy)_2Ir(L\text{-}pyr)]^{3+}$

$[M(phen)_3]^{2+}$ (M = Ru or Os)

$[(bpy)_2M(dpp)]^{2+}$ (M = Ru or Os)

$[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH\text{-}4\text{-}py)]^{2+}$

$[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH(CH_2)\text{-}4\text{-}py)]^{2+}$

$[\{(tpyPh(Ph_2P))_2Os\}]^{2+}$

$[(ppy)_2Ir(bpy)]^{+}$

$[(bpy)_2Ru(phenNHCO(COOHbpy))]^{2+}$

Eosin Y

$[Pt(CH_3Ph(tpy))(C_2Ph)]^{+}$

$[(tBu_2bpy)_2Ru(tpphz)]^{2+}$

$[(bpy)_2Ru(Mebpy(CH_2)_2C_6H_2(OCH_3)_2(CH_2)_2Mebpy)]^{2+}$

ZnTPP

Hydrogen via Direct Solar Production. Figure 1
Representative light absorbers used in solar hydrogen production schemes

Electron Relay Multicomponent systems rely on a suitable relay to transport electrons from an excited LA to the CAT. This electron relay must be able to accept electrons from the *LA. In supramolecular complexes, bridging ligands (BL) are used as electron relays and serve to covalently couple the individual components. A wide range of BLs has been used for this purpose, providing varied degrees of coupling between the LA and CAT and with varied acceptor orbital energetics [12]. Polyazine BLs are commonly

Hydrogen via Direct Solar Production. Table 1 Spectroscopic and redox properties of Ru^{II} and Os^{II} polyazine complexes

Complex	λ_{max}^{abs} (nm)	λ_{max}^{em} (nm)	τ (ns)	Ref.
$[Ru(bpy)_3]^{2+}$ [a]	449	605	850	[13]
$[Ru(phen)_3]^{2+}$ [a]	443	604	400	[13]
$[Ru(Ph_2phen)_3]^{2+}$ [b]	460	613	4,680	[23]
$[Os(bpy)_3]^{2+}$ [a]	640	723	20[b]	[24]
$[(bpy)_2Ru(dpp)]^{2+}$ [a]	470	675	240	[25, 26]
$[(phen)_2Ru(dpp)]^{2+}$ [a]	465	652	252	[27]
$[(bpy)_2Os(dpp)]^{2+}$ [a]	485	798	37	[28, 29]
$[Ru(tpy)_2]^{2+}$	476[c]	629[a]	0.25[b]	[30, 31]
$[Os(tpy)_2]^{2+}$	477[c]	718[c]	269[a]	[31]

[a] In acetonitrile solution at room temperature

[b] In aqueous solutions

[c] In ethanol–methanol (4/1)

used to couple molecular components within supramolecular assemblies. Polyazine BLs separated by short rigid spacers allow for significant electronic communication between components and directional electron or energy transfer [32–35]. $[(bpy)_2Ru(dpp)]^{2+}$ is a well-studied LA derived from $[Ru(bpy)_3]^{2+}$ with one of the bpy replaced by the polyazine BL, dpp (dpp = 2,3-bis(2-pyridyl)pyrazine) [25, 26, 33, 34]. The lowest lying excited state of $[Ru(bpy)_2(dpp)]^{2+}$ is a $Ru(d\pi) \rightarrow dpp(\pi^*)$ CT state, transferring charge from the ruthenium to the dpp ligand. This ^{1}MLCT absorption occurs at 470 nm with the ^{3}MLCT emission at 675 nm ($\tau = 210$ ns). The two remote nitrogen donor atoms on the dpp ligands provide points of attachment for a second metal center, allowing the construction of polymetallic complexes. The $dpp(\pi^*)$ acceptor orbital is stabilized when it is bridging two metals. In the homobimetallic system $[\{(bpy)_2Ru\}_2(dpp)]^{4+}$, a red shifted $Ru(d\pi) \rightarrow dpp(\pi^*)$ CT transition at 525 nm is observed with a shortened ^{3}MLCT lifetime of 80 ns [10, 25].

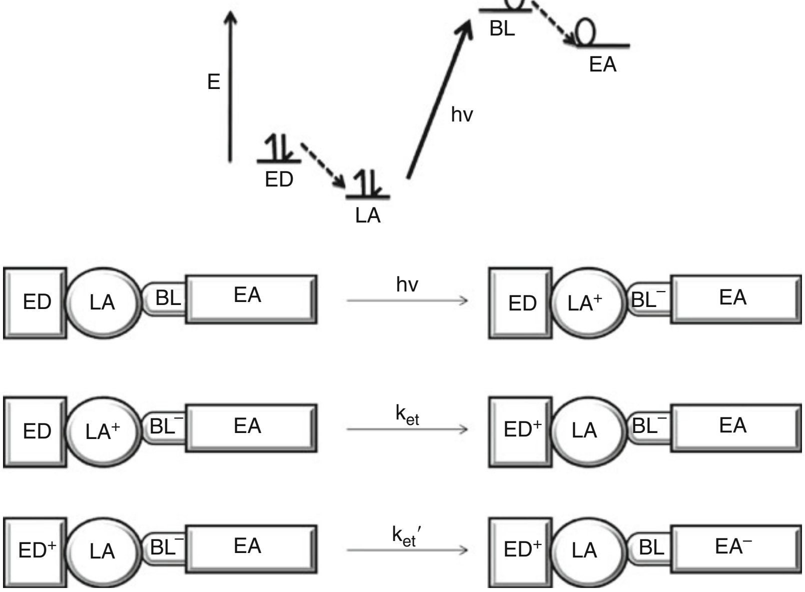
Catalysts Photochemical water reduction to produce hydrogen is an energetically uphill process involving bond dissociation and formation. Catalysts are used to promote this reaction. Most multicomponent systems are heterogeneous with considerable progress being made recently in homogeneous photocatalysis [19–21]. These systems typically use Pt, Pd, Rh, Co, or bioinspired diiron dithiolate (contains an Fe_2 core) catalysts. The non-precious metal-based Co and Fe systems are proposed as being economically viable. Homogeneous catalysis allows catalyst engineering to promote electron transfer from the LA and allows understanding of the mechanism of action through conventional solution-based analysis methods. Care must be taken in the selection of CATs to ensure that that they do not compete for light with the LA.

Turnover Number (TON) and Quantum Yield (Φ) The observed TON and Φ of a reaction are key parameters describing catalytic system efficiency and stability. The TON relative to the CAT is defined as the ratio of amount of hydrogen produced (n H_2) to the initial amount of CAT present (n_o CAT), Eq. 16:

$$\mathrm{TON} = \frac{n\,\mathrm{H_2}}{n_o\,\mathrm{CAT}} \tag{17}$$

Photoinitiated Charge Separation Solar energy conversion schemes that proceed via photoreduction of substrates are initiated with photoinduced charge separation that generates a potential at the molecular level. A simplified supramolecular device for photoinitiated charge separation is represented in Fig. 2. Optical excitation, followed by intramolecular electron transfer affords a charge separated state consisting of a positively charged electron donor, ED^+, and negatively charged electron acceptor, EA^-. The photogenerated charge separated (CS) state should persist for enough time to drive useful chemistry during the excited state lifetime. In supramolecular systems, the duration of charge separation can be modulated by fine tuning the orbital overlap between the ED and LA through modification of distance between these two subunits. The efficiency of charge separation is also key to functioning and often factors enhancing lifetime reduce quantum efficiency.

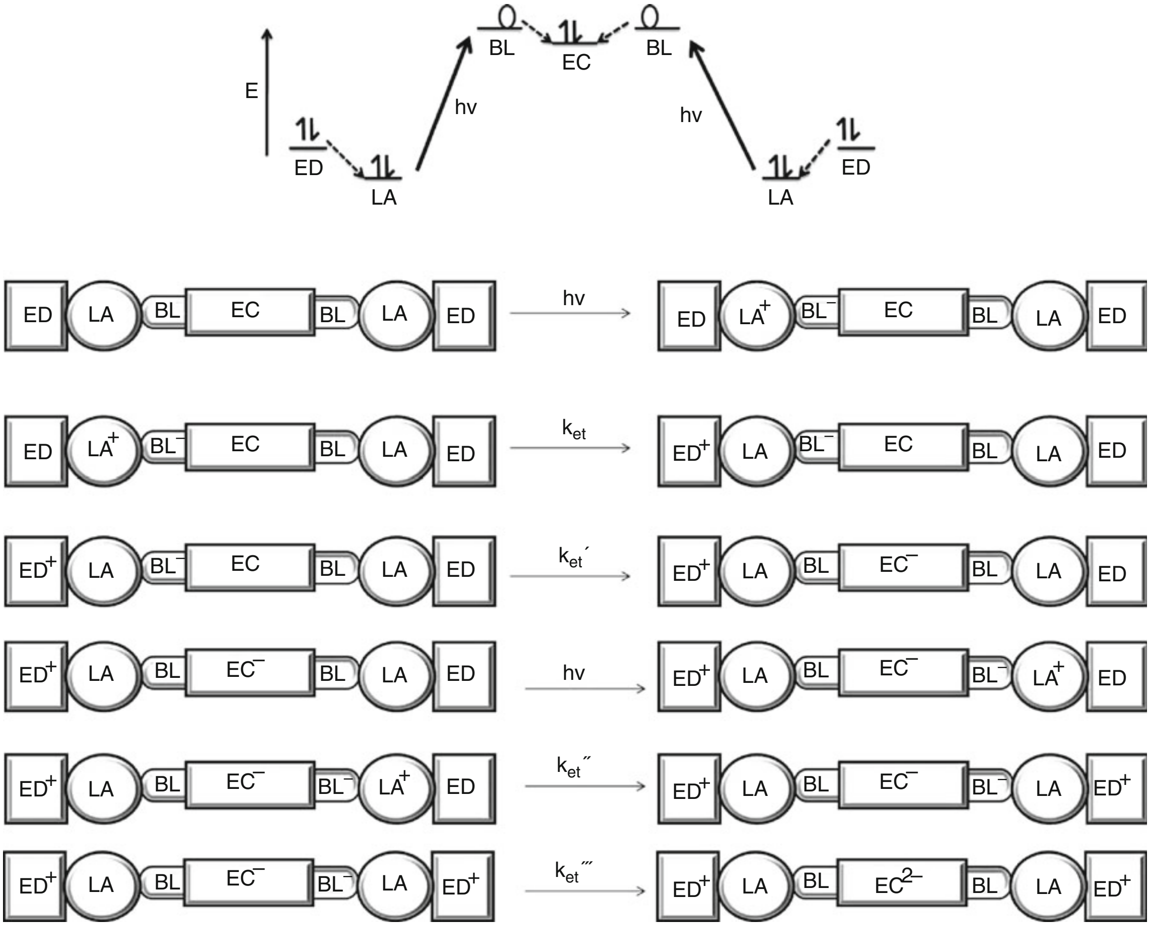
Photoinitiated Electron Collection Photoinitiated electron collection (PEC) is a process by which at least two electrons are collected on a central site



Hydrogen via Direct Solar Production. Figure 2
Photoinduced charge separation in a ED-LA-BL-EA molecular device (*ED* electron donor, *LA* light absorber, *EA* electron acceptor)

following optical excitation. For PEC, supramolecular complexes can be designed so that two ED-LA units are connected to a single EA that can accept at least two electrons (an electron collector, EC) to give a ED-LA-BL-EC-BL-LA-ED structural motif. Following multiple optical excitations, this system collects two electrons at the EC creating two ED^+, Fig. 3. PEC is of considerable importance in solar fuel production as multielectron reduction of the substrate is required in many fuel producing reactions. In the important monograph by Balzani in 1987, the design considerations for functioning photoinitiated electron collectors were discussed [8]. Since then many supramolecular systems that couple multiple LAs have been constructed, but only a few functioning PEC systems exist.

Photoinitiated Electron Collectors that Collect Electrons on Ligand Orbitals Photoinitiated electron collectors are devices that allow the use of light energy to collect the reducing equivalents needed to produce fuels. Molecular-based PEC systems are engineered so that the subunits are connected in an order with orbital energetics that allow multiple electrons to be collected on a single site following light activation. Current systems that undergo PEC have LA units connected to EC through BLs that function as electron relays. The first functioning photoinitiated electron collector is $[\{(bpy)_2Ru(dpb)\}_2IrCl_2](PF_6)_5$ (dpb = 2,3-bis(2-pyridyl)benzoquinoxaline) which collects two electrons on the dpb(π^*) orbitals upon two sequential excitations and electron transfer events, Fig. 4 [36]. Electron transfer from the dpb(π^*) orbitals to the Ir-based orbitals is energetically uphill. Homobimetallic ruthenium systems, $[(phen)_2Ru(tatpq)Ru(phen)_2]^{4+}$ (phen = 1,10-Phenanthroline, tatpq = 9,11,20,22-tetraazatetrapyrido[3,2-a:2′3′-c:3″,2″-1:2‴,3‴-n]pentacene-10,21-quinone) and $[(phen)_2Ru(tatpp)Ru(phen)_2]^{4+}$ (tatpp = 9,11,20,22- tetraazatetrapyrido[3,2-a:2′3′-c:3″,2″-1:2‴,3‴-n]pentacene) are the next PECs studied with multiple electrons also being collected on the BL(π^*) orbitals, Fig. 4 [37, 38]. Although these initial systems undergo PEC, they are not known to reduce substrates. The monometallic system, $[Ru(bpy)_2(pbn)]^{2+}$ (pbn = 2-(2-pyridyl)benzo[b]-1,5-naphthyridine), couples a Ru-based LA to an NAD^+ (NAD^+ = nicotinamide adenine dinucleotide) model ligand, Fig. 4. The NAD^+ model ligand acts as a reservoir for two electrons and uses



Hydrogen via Direct Solar Production. Figure 3
Photoinduced electron collection in a ED-LA-EA-LA-ED molecular device (*ED* electron donor, *LA* light absorber, *EA* electron acceptor)

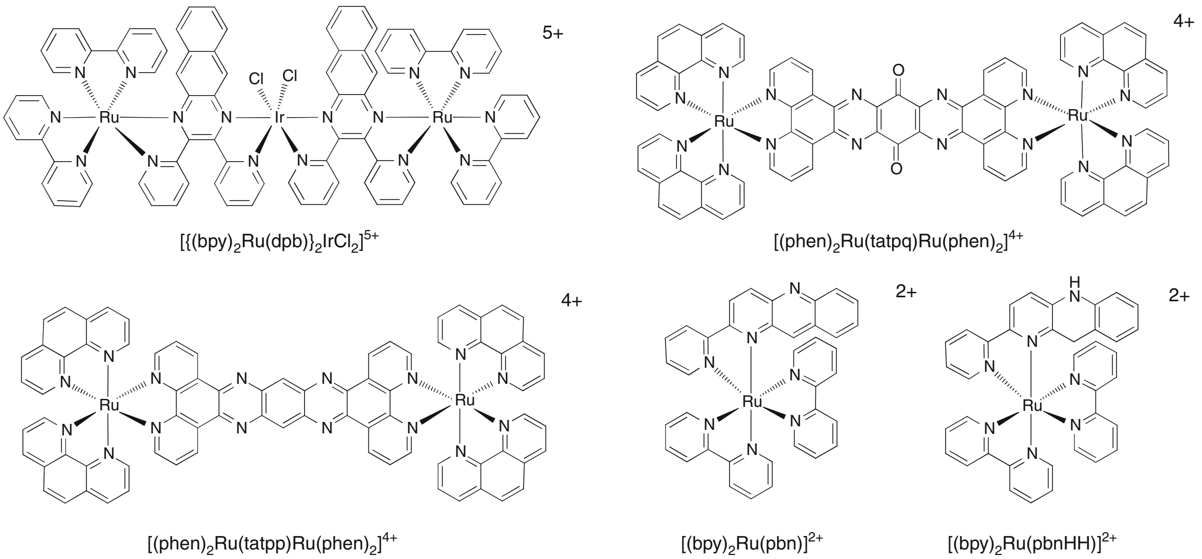
proton coupled electron transfer. Photolysis of a solution containing $[Ru(bpy)_2(pbn)]^{2+}$ and triethylamine (Et_3N) leads to the formation of NADH like $[Ru(bpy)_2(pbnHH)]^{2+}$, a potential hydride donor for reduction of organic substrates [39, 40].

Solar Hydrogen Photocatalysis

This entry will focus on the recent developments in molecular-based solar hydrogen photocatalysis using Fe, Co, Rh, Pt, and Pd-based CATs and Ru, Os, Pt, Re, and Ir-based LAs. Illustrative examples will be provided with discussions on system function, TON, and efficiency. The hydrogen TON, quantum yield, and system properties are summarized in Table 2.

Iron-Based System

Metabolism of protons in certain bacteria and algae are catalyzed by a certain class of metalloenzymes coined hydrogenases. Iron hydrogenases are important in proton reduction producing hydrogen gas. Recent research efforts focus on biomimetic diiron systems containing an Fe_2S_2 core, structurally similar to the iron hydrogenase active sites, for solar hydrogen photocatalysis [19–21, 41]. Interest in these systems has been inspired by the superior catalytic activity of hydrogenases which can reach 6,000–9,000 molecules of hydrogen s^{-1} [42]. Most Fe_2S_2 systems have been used in electrocatalytic proton reduction, requiring strongly acidic conditions and large overpotentials [43].



$[\{(bpy)_2Ru(dpb)\}_2IrCl_2]^{5+}$ $[(phen)_2Ru(tatpq)Ru(phen)_2]^{4+}$ $[(phen)_2Ru(tatpp)Ru(phen)_2]^{4+}$ $[(bpy)_2Ru(pbn)]^{2+}$ $[(bpy)_2Ru(pbnHH)]^{2+}$

Hydrogen via Direct Solar Production. Figure 4
Photoinitiated electron collectors that collect electrons on ligand orbitals

Supramolecular Iron-Based System The Ru-Fe_2S_2 dyads, $[(bpy)_2Ru(MebpyCONHPhCOO(Fe_2\{\mu\text{-}S_2(CH_2)_2CH\}(CO)_6))]^{2+}$, $[(tpy)Ru(tpyC_2Ph(Fe_2\{\mu\text{-}S_2(CH_2)_2\ N\}(CO)_6))]^{2+}$, $[(bpy)_2Ru(MebpyCONHPh(Fe_2\{\mu\text{-}S_2(CH_2)_2\ N\}(CO)_6))]^{2+}$, and $[(bpy)_2Ru(bpyC_2Ph(Ph_2)P(Fe_2\{\mu\text{-}S_2(CH_2)_3\}(CO)_5))]^{2+}$, in which a model iron hydrogenase active site is attached to a $[Ru(bpy)_3]^{2+}$ type chromophore, are supramolecular architectures constructed for possible application in solar hydrogen production, Fig. 5 [44–47]. These systems couple the favorable optical properties of the Ru-based chromophore to the catalytic properties of the diiron active site. It was anticipated that the optically populated ^{3}MLCT excited state would reduce the diiron species by two electrons and catalyze water reduction to produce hydrogen. It was found that electron transfer to the diiron dyads is thermodynamically unfavorable in this structural motif as the reduction potentials of the diiron units are more negative than the excited state oxidation potential of the ruthenium LAs. These Ru-Fe_2S_2 dyads are not solar hydrogen photocatalysts, but provide important design considerations for catalyst engineering. Component modification at the diiron center to facilitate electron transfer from the ^{3}MLCT excited state of the photosensitizer and/or photogeneration of a ^{3}MLCT excited state with a higher energy acceptor orbital have been suggested as a means to promote excited state electron transfer from the ruthenium LA to the diiron core [45–47].

A pyridine (py)-functionalized hydrogenase active site model complex bound to a zinc tetraphenylporphyrin [ZnTPP] LA, $[ZnTPP(pyCOO(CH_2)_2(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_6))]$, has been used in solar hydrogen production, Fig. 6 [48]. The photoinduced intramolecular electron transfer from the ^{1}MLCT state of the LA to the diiron unit has been established by fluorescence spectroscopy and laser flash photolysis technique. The weak interaction between the diiron center and the LA allows the two units to separate following intramolecular electron transfer, minimizing charge recombination through back electron transfer. The system can also reassemble to mediate intramolecular electron transfer.

Photolysis of ZnTPP, $[pyCOO(CH_2)_2(Fe_2\{\mu\text{-}S_2(CH_2)_2\ N\}(CO)_6)]$, and the electron donor 2-mercaptobenzoic acid in presence of CF_3COOH acid as the proton source in methylene chloride at $\lambda > 400$ nm resulted in photochemical hydrogen production with 0.16 TON based on $[pyCOO(CH_2)_2(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_6)]$ and 16 TON based on [ZnTPP]. The amount of hydrogen increases almost linearly in the first 40 min and reaches a plateau in 2 h. This decrease in rate of hydrogen production is attributed to CAT decomplexation through CO loss.

Hydrogen via Direct Solar Production. Table 2 Summary of hydrogen photocatalysis data

Solvent (electron donor)	λ (nm)	LA	LA-CAT	CAT	TON (time, h)	QY	Ref
CH_2Cl_2/CF_3COOH (2-mercaptobenzoic)	≥400		$[ZnTPP(pyCOO(CH_2)_2(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_6))]$		0.16 (Fe_2) (2)		[48]
					16 (LA) (2)		
CH_3CN/H_2O (ascorbic acid)	>400	$[Ru(bpy)_3]^{2+}$		$[Ph(CH_2)(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_5(P(N\text{-}pyr)_3))]$	4.3 (Fe_2) (3)		[51]
					86 (LA) (3)		
CH_3CN/H_2O (ascorbic acid)	>400	$[(bpy)_2Ru(Me_2bpy)]^{2+}$		$[Ph(CH_2)(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_5(P(N\text{-}pyr)_3))]$	3.9 (Fe_2) (3)		[51]
					76 (LA) (3)		
CH_3CN/H_2O (ascorbic acid)	>400	$[Ru(Me_2bpy)_3]^{2+}$		$[Ph(CH_2)(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_5(P(N\text{-}pyr)_3))]$	1.9 (Fe_2) (3)		[51]
					38 (LA) (3)		
CH_3CN/H_2O (ascorbic acid)	>400	$[Ru(bpy)_3]^{2+}$		$[Ph(CH_2)(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_6)]$	0.78 (Fe_2) (3)		[51]
CH_3CN/H_2O (ascorbic acid)	>400	$[Ru(bpy)_3]^{2+}$		$[Ph(CH_2)(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_4(P(N\text{-}pyr)_3)_2)]$	1.7 (Fe_2) (3)		[51]
					34 (LA) (3)		
DMF/H_2O (ascorbic acid)	455–850	$[Ru(bpy)_3]^{2+}$		$[Fe_2(\mu\text{-}Cl_2bdt)(CO)_6]$	200 (Fe_2) (2.5)	0.01	[52]
CH_3COCH_3/H_2O (Et_3N)	>400	$[(ppy)_2Ir(bpy)]^+$		$[Ph(CH_2)(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_5(P(N\text{-}pyr)_3))]$	466 (Fe_2) (8)		[53]
					660 (LA) (8)		
CH_3CN/H_2O (N,N′-Dimethylaniline)	470		$[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$		28 (CAT) (4)		[54]
					20 (CAT) (2)		
CH_3CN/H_2O (N,N′-Dimethylaniline)	470		$[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$		38 (CAT) (4)		[55–57]
					26 (CAT) (2)		
					3 (CAT) (4)[a]		
CH_3CN/H_2O (N,N′-Dimethylaniline)	470		$[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$		18 (CAT) (2)		[56]

CH_3CN/H_2O (triethylamine)	470		$[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$		4 (CAT) (2)		[56]
CH_3CN/H_2O (triethylamine)	470		$[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$		6 (CAT) (2)		[56]
CH_3CN/H_2O (triethylamine)	470		$[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$		3 (CAT) (2)		[56]
CH_3CN/H_2O (triethanolamine)	470		$[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$		1 (CAT) (2)		[56]
CF_3SOOH/H_2O (sodium ascorbate)	>380		$[\{(tpyPh(Ph_2P))_2Os\}RhCl(CO)]^{2+}$		36 (CAT) (18)		[58]
CF_3SOOH/H_2O (sodium ascorbate)	>380		$[\{(tpyPh(Ph_2P))_2Os\}RhCl_2]^{3+}$		87 (CAT) (18) 594 (CAT) (240)	0.007	[58]
CF_3SOOH/H_2O (sodium ascorbate)	>380	$[Os(tpy)_2]^{2+}$		$RhCl_3 \bullet 3H_2O$/dppe	24 (CAT) (18)		[58]
HCl/THF	300–400		$[Rh_2^{0,II}(dfpma)_3X_2(L)]$		27/h (CAT) (3)	0.01	[59]
H_2O (triethanolamine)	450	$[Ru(bpy)_3]^{2+}$, $[Rh(bpy)_3]^{3+}$		Pt	500 (LA) (240)	0.11^b	[11, 60]
THF/H_2O (triethanolamine)		$[Ir(f\text{-}mppy)_2(dtbbpy)]^+$		$[Rh(dtbbpy)_3]^{3+}$	5,000 (LA or CAT) (20–24)	0.34	[61]
CH_3COCH_3/Et_3NH^+ (triethylamine)^c	UV-VIS		$[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$		56 (Co) (4)		[62]
CH_3COCH_3/H_2O (triethylamine)^c	UV-VIS		$[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$		22 (Co) (4)		[62]
CH_3COCH_3/Et_3NH^+ (triethylamine)^c	>350		$[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$		103 (Co) (15)		[62]
CH_3COCH_3/Et_3NH^+ (triethylamine)^c	UV-VIS		$[(bpy)_2Ru(L\text{-}pyr)Co(dmgH)_2Cl]^{2+}$		17 (Co) (4)		[62]
CH_3COCH_3/Et_3NH^+ (triethylamine)^c	UV-VIS		$[(bpy)_2Ru(L\text{-}pyr)Co(dpgBF_2)_2(OH_2)]^{2+}$		12 (Co) (4)		[62]

Hydrogen via Direct Solar Production. Table 2 (Continued)

Solvent (electron donor)	λ (nm)	LA	LA-CAT	CAT	TON (time, h)	QY	Ref
CH_3COCH_3/Et_3NH^+ (triethylamine)[c]	>380		$[(Me_2phen)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$		9 (Co) (4)		[63]
CH_3COCH_3/Et_3NH^+ (triethylamine)			$[(ppy)_2Ir(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^+$		90 (Co) (4)		[63]
					210 (Co) (15)[e]		
CH_3COCH_3/Et_3NH^+ (triethylamine)[e]	>380	$[Ir(ppy)_2(phen)]^+$		$[Co(dmgBF_2)_2(OH_2)_2]$	165 (Co) (15)	0.11[d]	[63]
CH_3COCH_3/Et_3NH^+ (triethylamine)[e]	>380	$[Re(CO)_3Br(phen)]$		$[Co(dmgBF_2)_2(OH_2)_2]$	273 (Co) (15)	0.16[f]	[63]
CH_3COCH_3/Et_3NH^+ (triethylamine)[d]	>400		$[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$		38 (Co) (8)		[64]
CH_3COCH_3/Et_3NH^+ (triethylamine)[d]	>400		$[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH(CH_2)\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$		47 (Co) (8)		[64]
CH_3CN/H_2O (Eu^{2+})	450	$[Ru(bpy)_3]^{2+}$		$[Co^{II}L]$	10 (LA)		[65]
					1 (Co)		
CH_3CN/H_2O (triethanolamine)	465	$[Ir(F\text{-}mppy)_2(bpy)]^+$		$[Co(bpy)_3]^{2+}$	920 (LA)		[66]
					18 (Co)		
CH_3CN/H_2O (triethanolamine)	>410	$[Pt(ttpy)(C \equiv CPh)]ClO_4$		$[Co(dmgH)_2pyCl]$	2,100 (10)		[67]

CH_3CN/H_2O (triethanolamine)	>450	Eosin Y		$[Co(dmgH)_2pyCl]^{2+}$	900 (LA) (14)[g]	0.04[h]	[68]
$DMF/H_2O/AcOH$ (triethanolamine)	>400	$[ReBr(CO)_3bpy)]$		$[Co(dmgH)_2]$	150 (LA) (9)	0.26	[69]
H_2O/acetate buffer (EDTA)	>350		$[(bpy)_2Ru(phenNHCO(COOHbpy))PtCl_2]^{2+}$		5 (CAT) (10)	0.01	[70]
CH_3CN/H_2O (N,N′-Dimethylaniline)	470		$[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$		115 (CAT) (5)		[71]
CH_3CN/H_2O (N,N′-Dimethylaniline)	530		$[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$		40 (CAT) (3)	0.01	[72]
H_2O/MES (EDTA)	350–550		[PtCl(tpy)]Cl		3 (LA or CAT) (7)	0.02	[73]
CH_3CN/triethylamine (triethylamine)	470		$[(^tBu_2bpy)_2Ru(tpphz)PdCl_2]^{2+}$		56 (CAT) (30)		[74]

[a] In water/phosphate buffer using TEOA ED (Ref. [57])

[b] Ref. [11]

[c] 100-fold excess of both Et_3NH^+ and Et_3N were used

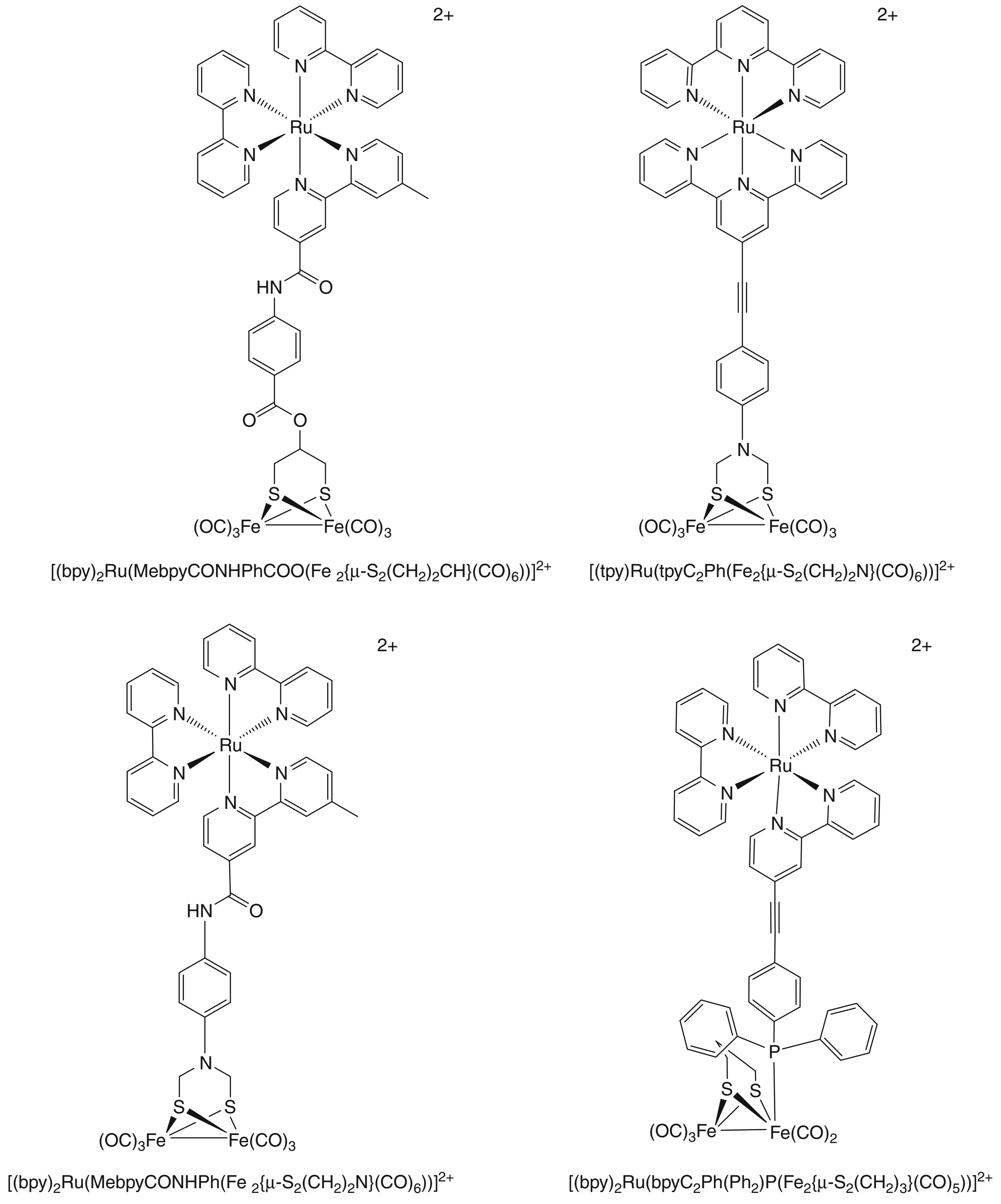
[d] 300-fold excess of both Et_3NH^+ and Et_3N were used

[e] 600-fold excess of both Et_3NH^+ and Et_3N were used

[f] $\lambda = 412$ nm, irradiation time 2 h

[g] dmgH was added to the system

[h] $\lambda = 520$ nm, irradiation time 2 h



$[(bpy)_2Ru(MebpyCONHPhCOO(Fe_2\{\mu\text{-}S_2(CH_2)_2CH\}(CO)_6))]^{2+}$

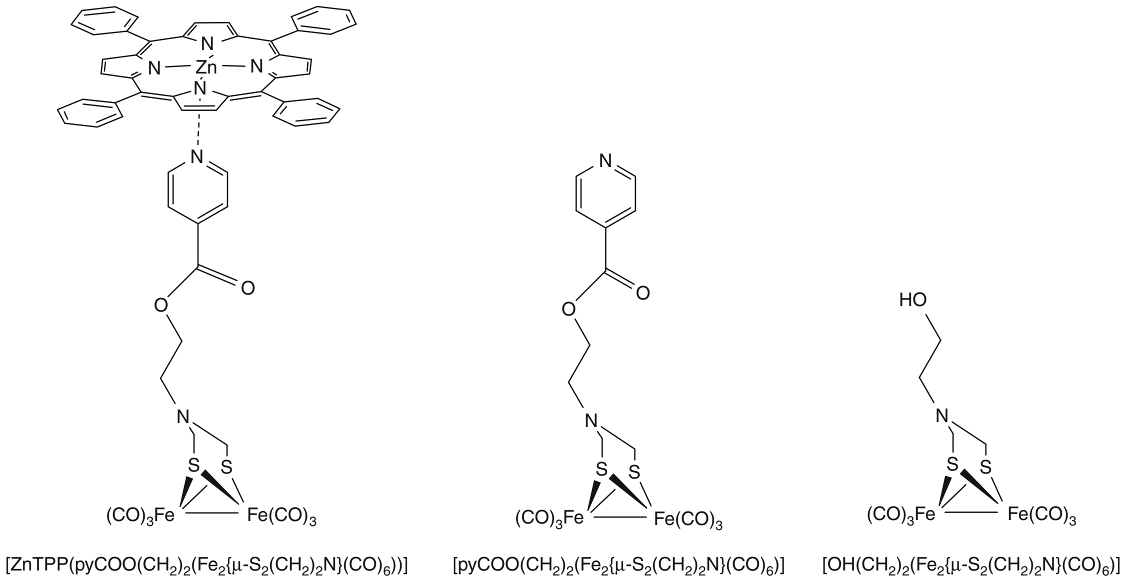
$[(tpy)Ru(tpyC_2Ph(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_6))]^{2+}$

$[(bpy)_2Ru(MebpyCONHPh(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_6))]^{2+}$

$[(bpy)_2Ru(bpyC_2Ph(Ph_2)P(Fe_2\{\mu\text{-}S_2(CH_2)_3\}(CO)_5))]^{2+}$

Hydrogen via Direct Solar Production. Figure 5
Systems that couple a ruthenium chromophore to an iron-based catalyst

Hydrogen production does not occur when $[pyCOO(CH_2)_2(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_6)]$ is replaced by $[OH(CH_2)_2(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_6)]$. This suggests that assembly of the system and the intramolecular electron transfer events are crucial for solar hydrogen catalysis as $[OH(CH_2)_2(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_6)]$ does not

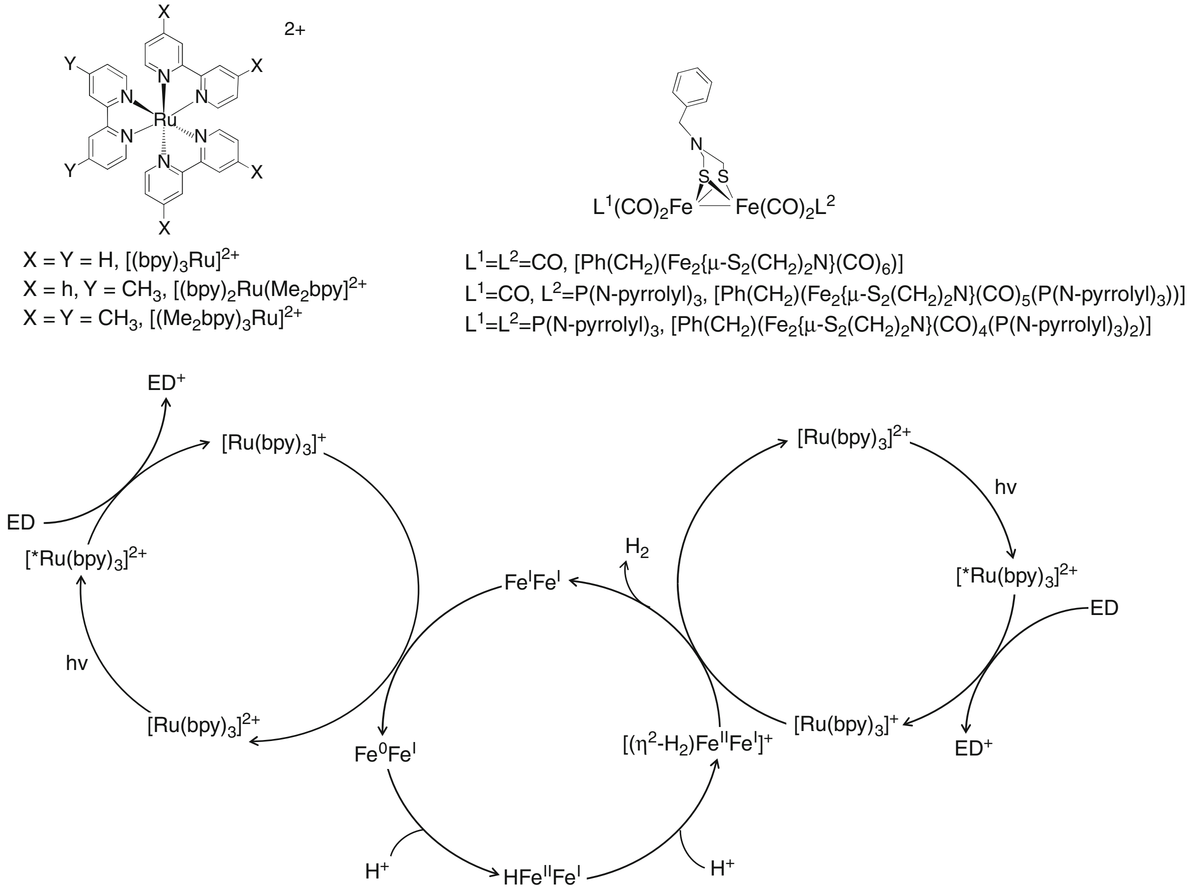

[ZnTPP(pyCOO(CH2)2(Fe2{μ-S2(CH2)2N}(CO)6))]
[pyCOO(CH2)2(Fe2{μ-S2(CH2)2N}(CO)6)]
[OH(CH2)2(Fe2{μ-S2(CH2)2N}(CO)6)]

Hydrogen via Direct Solar Production. Figure 6
Representations of ZnFe self-assembling system and the FeFe subunits

coordinate with ZnTPP and hydrogen production does not occur. It is proposed that the singlet excited state intramolecular electron transfer forms the charge separated state [ZnTPP$^{\cdot+}$pyCOO(CH$_2$)$_2$(Fe$_2${μ-S$_2$(CH$_2$)$_2$N}(CO)$_6$)$^{\cdot-}$], which undergoes dissociation to form [ZnTPP$^{\cdot+}$] and [pyCOO(CH$_2$)$_2$(Fe$_2${μ-S$_2$(CH$_2$)$_2$N}(CO)$_6$)]$^{\cdot-}$. [ZnTPP$^{\cdot+}$] is reductively quenched by the ED, while [pyCOO(CH$_2$)$_2$(Fe$_2${μ-S$_2$(CH$_2$)$_2$N}(CO)$_6$)]$^{\cdot-}$ is protonated. The protonated radical anion facilitates the second electron transfer event, which is followed by further protonation and hydrogen release. Further protonation of the protonated radical anion prior to the second electron transfer has been observed in electrochemical hydrogen production catalyzed by Fe$_2$S$_2$ hydrogenases [49]. This is the first photoactive hydrogenase model system that incorporates the self-assembling principle for photochemical hydrogen production. This system is interesting as both CAT and LA are based on bioinspired materials.

In photosynthesis, the light-dependent reactions are carried out by two photosystems, photosystem I (PS I) and photosystem II (PS II). PS I catalyzes the optically induced electron transfer from reduced cytochrome c to oxidized ferredoxin that are directly used to reduce nutrients or used to produce NADH. Systems that couple PS I to active hydrogen producing CATs are known. A recent review describes this technology employing molecular wires to connect PS I directly to CAT including Pt nanoparticles or the distal [4Fe-4S] of a Fe$_2$ or NiFe hydrogenase [50].

Multicomponent Iron-Based Systems A few multicomponent systems have been reported that use biomimetic iron hydrogenase model complexes as CATs in solar hydrogen production schemes. Multicomponent systems are solution systems with multiple molecules interacting through diffusion to allow individual molecules to perform the acts associated with solar hydrogen production. The photocatalytic system typically consists of three main components: LAs, CATs, and EDs functioning through intermolecular interactions in solution. A photoactive multicomponent system for solar hydrogen production using biomimetic iron hydrogenase model complexes was first reported in 2008 [51]. This system uses [Ru(bpy)$_3$]$^{2+}$ type LAs, Fe$_2$S$_2$ CATs, and ascorbic acid as the ED irradiating at $\lambda > 400$ nm, Fig. 7. The system consisting of [Ru(bpy)$_3$]$^{2+}$ LA, [Ph(CH$_2$)(Fe$_2${μ-S$_2$(CH$_2$)$_2$ N}(CO)$_6$)] CAT containing all CO ligands, and ascorbic acid in CH$_3$CN/H$_2$O has the lowest activity attributed to decomplexation of the CO from the CAT. One or two of the CO ligands on the Fe$_2$S$_2$


2+
X
Y
N
Ru
X = Y = H, [(bpy)3Ru]2+
X = h, Y = CH3, [(bpy)2Ru(Me2bpy]2+
X = Y = CH3, [(Me2bpy)3Ru]2+
N
S S
L1(CO)2Fe Fe(CO)2L2
L1=L2=CO, [Ph(CH2)(Fe2{μ-S2(CH2)2N}(CO)6)]
L1=CO, L2=P(N-pyrrolyl)3, [Ph(CH2)(Fe2{μ-S2(CH2)2N}(CO)5(P(N-pyrrolyl)3))]
L1=L2=P(N-pyrrolyl)3, [Ph(CH2)(Fe2{μ-S2(CH2)2N}(CO)4(P(N-pyrrolyl)3)2)]
ED+
[Ru(bpy)3]+
ED
[*Ru(bpy)3]2+
hv
[Ru(bpy)3]2+
Fe0FeI
FeIFeI
H2
[(η2-H2)FeIIFeI]+
H+
HFeIIFeI
H+
[Ru(bpy)3]2+
hv
[*Ru(bpy)3]2+
ED
[Ru(bpy)3]+
ED+

Hydrogen via Direct Solar Production. Figure 7
Homogeneous three-component system for solar hydrogen production catalyzed by biomimetic diiron hydrogenase catalyst

unit were replaced by a phosphine-based ligand to improve the photostability and catalytic activity of the system. The photosystem consisting of the P(pyr)$_3$-monosubstituted CAT (Ph = phenyl, P(pyr)$_3$ = tris (N-pyrrolyl)phosphine), [Ph(CH$_2$)(Fe$_2${μ-S$_2$(CH$_2$)$_2$ N}(CO)$_5$(P(N-pyr)$_3$))], showed greater photostability and catalytic activity albeit low turnovers, 4.3 TON relative to CAT and 86 TON relative to the LA. The photosystem consisting of [Ru(bpy)$_3$]$^{2+}$ showed the greatest hydrogen production relative to [(bpy)$_2$Ru (Me$_2$bpy)]$^{2+}$ or [(Me$_2$bpy)$_3$Ru]$^{2+}$ LAs. The decline in the hydrogen production when [(bpy)$_2$Ru(Me$_2$bpy)]$^{2+}$ or [(Me$_2$bpy)$_3$Ru]$^{2+}$ is used is attributed to inefficient reductive quenching of the excited LA by the ascorbate anion. Methyl substitution renders the excited state reduction potential of the LA more negative making reductive quenching by ascorbate anion less thermodynamically favorable.

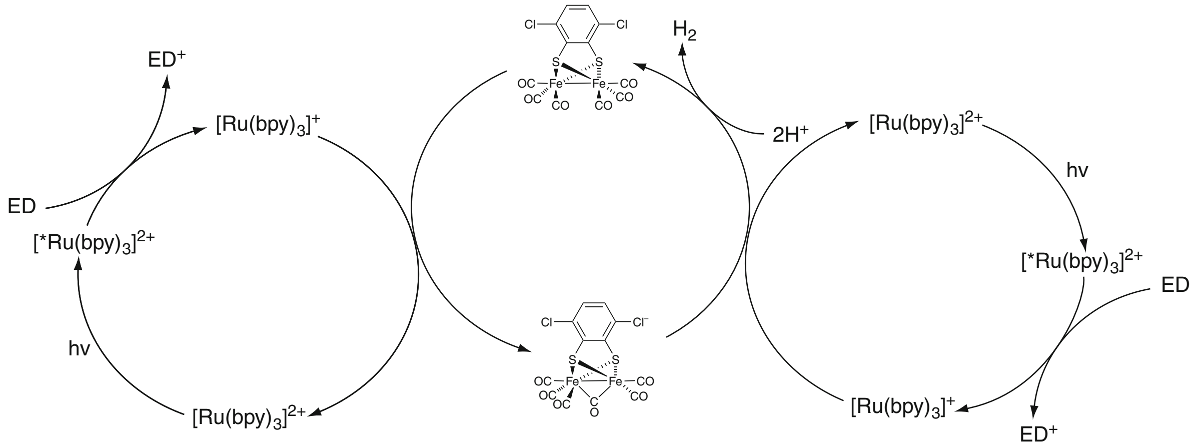
It is proposed that the excited [*Ru(bpy)$_3$]$^{2+}$ LA is reductively quenched by the electron donor to form [Ru(bpy)$_3$]$^+$. Intermolecular electron transfer to the diiron unit is thermodynamically feasible and a Fe^{0}Fe^I species is formed. Based on the redox potentials it is established that further reduction of the CAT is thermodynamically unfavorable. Protonation of the Fe^{0}Fe^I species to form HFe^{II}Fe^I followed by further reduction HFe^{I}Fe^I is likely. One proposal is that hydrogen production occurs through protonation of HFeIFeI with regeneration of the CAT. An alternative mechanism is the second protonation of HFeIIFeI to generate [(η^2-H$_2$)Fe^{II}Fe^I]$^+$ that is capable of being reduced to generate hydrogen and regenerate the CAT, Fig. 7.

Higher TON for photochemical hydrogen production has been achieved using $[Fe_2(\mu\text{-}Cl_2bdt)(CO)_6]$ (Cl_2bdt = 3,6-dichlorobenzene-1,2-dithiolate) as the CAT [52]. $[Fe_2(\mu\text{-}Cl_2bdt)(CO)_6]$ is a known electrocatalyst for hydrogen production, allowing electrochemical reduction of protons at low potentials [75]. This complex is unusual displaying electrochemical reversibility. Electrochemical reversibility in these systems could be particularly relevant for hydrogen photocatalysis as proton reduction occurs by consecutive one electron reduction steps. Photolysis of a system consisting of $[Ru(bpy)_3]^{2+}$, $[Fe_2(\mu\text{-}Cl_2bdt)(CO)_6]$, ($[Ru(bpy)_3]^{2+}/[Fe_2(\mu\text{-}Cl_2bdt)(CO)_6]$ 10:1), and ascorbic acid (pH = 5.5) in DMF/H_2O at λ = 455–850 nm provides over 200 TON relative to the CAT in 2.5 h at which time photocatalysis discontinues attributed to CAT deactivation. The maximum TON frequency is 2.7 H_2 CAT^{-1} min^{-1}. The Φ of hydrogen production is 0.01 per absorbed photon. A $\Phi = 0.014$ with a frequency of 3.7 H_2 CAT^{-1} min^{-1} has also been obtained at pH = 6.4, but with lower TON and system stability. The proposed mechanism of action is represented in Fig. 8. Transient absorption and emission spectral data have established the elementary steps of the catalytic cycle to involve the reductive quenching of the excited LA by ascorbic acid to form $[Ru(bpy)_3]^+$ and reoxidation of $[Ru(bpy)_3]^+$ by the diiron unit.

Photochemical hydrogen production in a more basic medium has been investigated using the three-component system using $[(ppy)_2Ir(bpy)]^+$ LA (ppy = phenylpyridine), $[Ph(CH_2)(Fe_2\{\mu\text{-}S_2(CH_2)_2N\}(CO)_5(P(N\text{-}pyr)_3))]$ CAT, and Et_3N electron donor in CH_3COCH_3/H_2O irradiating at $\lambda > 400$ nm [53]. Turnover numbers of 466 relative to the CAT and 660 relative to the LA in 8 h has been achieved in this system. The impact of solvent, ED, LA, and CAT concentrations on solar hydrogen production has been investigated to determine the optimized concentrations to achieve the high TON reported. Unlike in the previous diiron systems, system deactivation occurs through decomposition of the LA, as re-addition of the LA showed total system recovery. The mechanism of catalyst function under basic conditions is unclear but it is proposed to occur through a Fe^0Fe^I species which is formed through reductive quenching of the LA by Et_3N followed by reduction of the Fe^IFe^I form of the CAT.

Rh-Based Supramolecular and Multicomponent Systems

Tris-(2,2′-bipyridine)Rh^{III}, $[Rh(bpy)_3]^{3+}$, is known to undergo excited state electron transfer of $[^*Ru(bpy)_3]^{2+}$ [76]. The one electron reduction product, $[Rh(bpy)_3]^{2+}$,



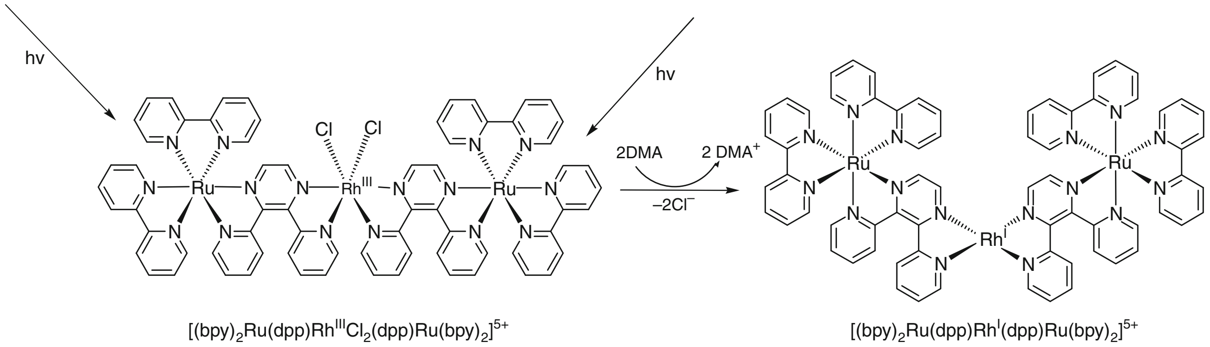
Hydrogen via Direct Solar Production. Figure 8
Proposed mechanism for photochemical hydrogen generation using $[Ru(bpy)_3]^{2+}$ LA, $[Fe_2(\mu\text{-}Cl_2bdt)(CO)_6]$ CAT, and ascorbic acid in DMF/H_2O

is kinetically unstable and rapidly disproportionate to form $[Rh(bpy)_3]^+$ and $[Rh(bpy)_3]^{3+}$. Rh-based systems are used as the electron relays in solar hydrogen production schemes.

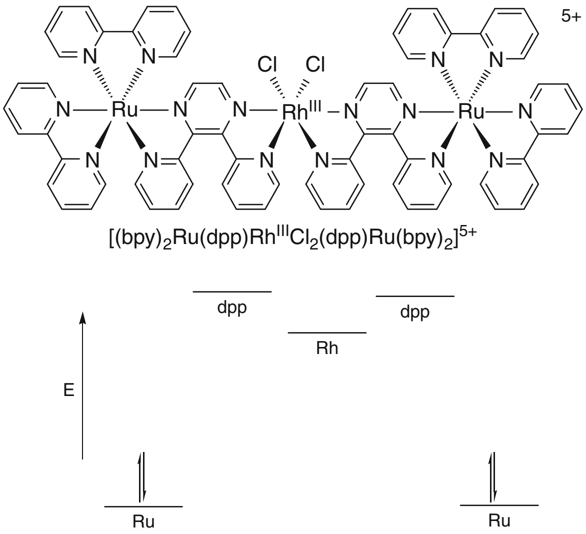
Systematic Studies Using Rh-Based Supramolecular Photocatalysts for Solar Hydrogen Production The coupling of Rh centers to Ru light absorbers in supramolecular frameworks is reported [77–81]. The majority of this work focuses on the study of chelated complexes of Rh^{III} coupled to Ru polyazine charge-transfer complexes. In these studies the Rh acts as an intramolecular electron acceptor for the MLCT excited state of the Ru LA. In many of the systems, the Rh centers are bound to only chelating ligands, which can limit the chemistry that can follow intramolecular electron transfer. Brewer et al. reported the supramolecular complex $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$, the first photoinitiated electron collector that collects multiple electrons on a metal capable of photocatalyzing hydrogen production from water [54–57, 82, 83]. This supramolecular system incorporates two $[(bpy)_2Ru(dpp)]^{2+}$ LAs and a single Rh EC allowing for multiple electron reduction. This system is unique from the previously reported systems as the Rh center is bound to monodentate Cl^- ligands that can be lost following photoreduction. This allows the metal to become coordinatively unsaturated, allowing complexation and/or reaction with substrates. The redox properties of $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$ demonstrate that the Ru LAs are electronically isolated, important to PEC and multielectron photochemistry.

Brewer et al. subsequently reported other systems incorporating a similar architecture that are photoinitiated electron collectors catalyzing solar hydrogen production. These systems also couple two Ru^{II} or Os^{II} LAs to a single Rh^{III} acceptor. The supramolecular assemblies provide a LA-BL-RhX_2-BL-LA structural motif (LA = Ru^{II} or Os^{II} polyazine LA, X = Cl^- or Br^-, BL = dpp). Visible light irradiation affords PEC resulting in Rh reduction and loss of the labile ligands, Fig. 9. Studies have established $[\{(bpy)_2Ru(dpp)\}_2RhX_2]^{5+}$ (X = Cl^- or Br^-) and $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$ as photochemical molecular devices for electron collection at a metal center and are photocatalysts for solar hydrogen production from water with a hydrogen yield of $\Phi \approx 0.01$ [54, 56, 82]. Modification of the LA to Os and/or the TL to tpy and a Cl destabilizes the LA metal-dπ orbitals providing systems that absorb light at lower energy. The complexes $[\{(bpy)_2Os(dpp)\}_2RhCl_2]^{5+}$, $[\{(tpy)OsCl(dpp)\}_2RhCl_2]^{3+}$, and $[\{(tpy)RuCl(dpp)\}_2RhCl_2]^{3+}$ have been studied and their photocatalytic activity evaluated [56]. Studies show that $[\{(bpy)_2Os(dpp)\}_2RhCl_2]^{5+}$ and $[\{(tpy)RuCl(dpp)\}_2RhCl_2]^{3+}$ are solar hydrogen photocatalysts but provide lower activity consistent with significantly lower driving forces for excited state reductive quenching by the ED. Factors that impact photocatalytic efficiency have been evaluated by studying the redox and photophysical properties of these systems.

Rhodium-Based Supramolecular Photocatalysts: Redox Properties The electrochemical analysis of the supramolecular systems provides meaningful


hv
hv
Cl
Cl
N
Ru
Rh^III
2DMA
2 DMA+
–2Cl–
Rh^I
[(bpy)2Ru(dpp)Rh^IIICl2(dpp)Ru(bpy)2]5+
[(bpy)2Ru(dpp)Rh^I(dpp)Ru(bpy)2]5+

Hydrogen via Direct Solar Production. Figure 9
Photoinitiated electron collection on a metal center



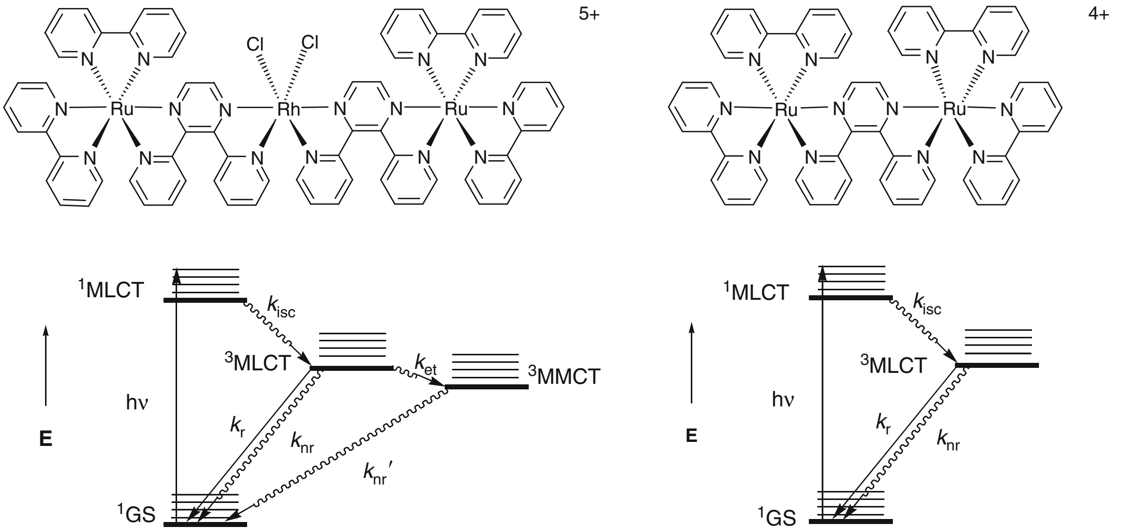
Hydrogen via Direct Solar Production. Figure 10
Orbital energy diagram of Rh-centered photoinitiated electron collection of the form LA-BL-RhX$_2$-BL-LA (LA = bpy or phen, BL = dpp, X = Cl or Br)

information to understand excited state electron transfer and photocatalyst function. Electrochemical analysis provides a good estimation of the energy of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), allowing the study of the orbital energetics. Electrochemically the trimetallic complexes of the general formula $[\{(TL)_2\,M(dpp)\}_2RhX_2]^{5+}$ (TL = bpy or phen, M = Ru or Os, X = Cl or Br) and $[\{(tpy)MCl(dpp)\}_2RhCl_2]^{3+}$ (M = Ru or Os) demonstrate reversible metal-based oxidations and irreversible Rh- and reversible ligand-based reductions [56]. All the trimetallics show overlapping Ru$^{II/III}$ or Os$^{II/III}$ based oxidations for the two terminal ligands, indicating minimal electronic interaction between the two terminal metal LA subunits. The Ru centers of the RuRu bimetallic systems $[\{(bpy)_2Ru\}_2(dpp)]^{4+}$ or $[\{(phen)_2Ru\}_2(dpp)]^{4+}$ are electronically coupled showing two Ru$^{II/III}$ oxidations at *ca.* +1.47 and +1.66 V vs. SCE [84]. These bimetallic systems do not display photoinitiated electron collection and/or reduction of substrates. $[\{(bpy)_2Ru(dpp)\}_2RhX_2]^{5+}$ systems (X = Cl or Br) or $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$ show overlapping oxidative couples at *ca.* 1.60 V vs. SCE. Rh^{III} centers coordinated to polyazine ligands typically display irreversible reductions corresponding to Rh$^{III/II/I}$ processes [85]. The Rh$^{III/II/I}$ reductions are followed by the loss of halides as observed for $[Rh(bpy)_2Cl_2]^+$. In the trimetallic systems, the Rh$^{III/II/I}$ reduction is followed by two reversible dpp$^{0/-}$ reductions. The halides on the Rh^{III} center impact the redox properties as evidenced by a slightly more positive Rh$^{III/II/I}$ reduction, −0.36 V vs. SCE for $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$ relative to $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$, −0.40 V vs. SCE [54–56]. This positive shift in the Rh$^{III/II/I}$ couple is consistent Br$^-$ being a weaker σ donor than Cl$^-$. The Os$^{II/III}$ couple occurs at 1.17 V vs. SCE in $[\{(bpy)_2Os(dpp)\}_2RhCl_2]^{5+}$. The more positive oxidation potential is a result of more destabilized Os(dπ) orbitals relative to the Ru (dπ) orbitals in the Ru-based systems. The Ru$^{II/III}$ oxidation of $[\{(tpy)RuCl(dpp)\}_2RhCl_2]^{3+}$ occurs at 1.09 V vs. SCE, more positive than the Ru$^{II/III}$ oxidations of the analogous bpy-based systems, and consistent with a more electron rich ruthenium center due to Cl$^-$ coordination in place of a pyridine ring. The Os$^{II/III}$ couple in $[\{(tpy)OsCl(dpp)\}_2RhCl_2]^{3+}$ occurs at 0.81 V vs. SCE, consistent with even more destabilized Os(dπ) orbitals. The reductive electrochemistry of $[\{(tpy)MCl(dpp)\}_2RhCl_2]^{3+}$ shows irreversible Rh$^{III/II/I}$ reductions at −0.51 and −0.55 V vs. SCE (M = Ru or Os), respectively, followed by two reversible dpp$^{0/-}$ couples. The electrochemistry of the trimetallic supramolecular complexes predict Ru(dπ)- or Os(dπ)-HOMOs with energy tuned by the TL or LA metal and Rh(dσ*)-based LUMOs with energy tuned by the halides bound to Rh. The electrochemistry predicts a lowest lying metal-to-metal charge-transfer state (MMCT) with an oxidized Ru and a reduced Rh, Fig. 10.

Rhodium-Based Supramolecular Photocatalysts: Spectroscopic Properties The optical properties of the trimetallic supramolecular complexes are evaluated by electronic absorption spectroscopy [54–56]. The trimetallic supramolecular assemblies are efficient light absorbers throughout the UV and visible regions of the spectrum with transitions characteristic of each subunit of the LA-BL unit. The electronic absorption spectra of $[\{(bpy)_2Ru(dpp)\}_2RhX_2]^{5+}$ (X = Cl or Br) and $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$ exhibit intense π → π* TL and dpp transitions in the UV, Ru(dπ) → TL (π*) (TL = bpy or phen) CT transitions between 410 and 420 nm, and Ru(dπ) → dpp(π*) CT transitions at

$\lambda_{max}{}^{abs}$ = 520 nm ($\varepsilon = 2.6 \times 10^4\ M^{-1}\ cm^{-1}$) [56]. The lowest energy Ru(dπ) → dpp(π*) CT transition for $[\{(bpy)2Ru\}_2(dpp)]^{4+}$ or $[\{(phen)_2Ru\}_2(dpp)]^{4+}$ occurs at $\lambda_{max}{}^{abs}$ = 526 nm demonstrating that the Ru (dp) → dpp(π*) CT are dominating in this region. This is verified by the nearly identical electronic absorption spectra of $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$ and $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$ demonstrating that the halide on the Rh does not significantly impact the optical properties. The electronic absorption spectra for the tpy-based systems are similar to the bpy analogs showing changes consistent with the destabilized metal orbitals showing Ru(dπ) → dpp(π*) CT transitions at $\lambda_{max}{}^{abs}$ = 540 nm. For the Os analogs, $[\{(bpy)_2Os(dpp)\}_2RhCl_2]^{5+}$ and $[\{(tpy)OsCl(dpp)\}_2RhCl_2]^{3+}$, the Os(dπ) → dpp(π*) CT transitions occur at slightly lower energies relative to the ruthenium-based systems, consistent with the destabilized Os(dπ) orbitals. The Os systems show pronounced intensity of the ^{3}MLCT absorption in the low energy region due to a high degree of spin-orbit coupling.

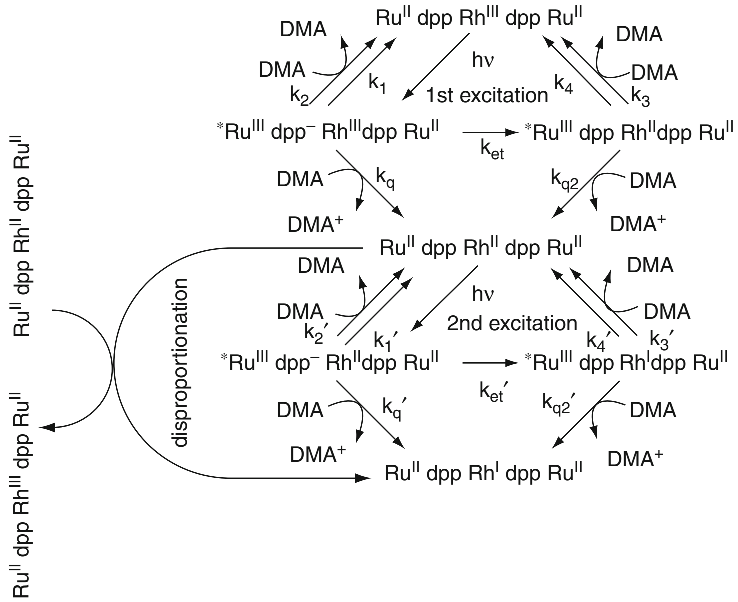
Rhodium-Based Supramolecular Photocatalysts: Photophysics Emission spectroscopy is used to probe intramolecular electron transfer within a supramolecular assembly possessing emissive MLCT light absorbers. The Ru-based $[\{(TL)_2Ru(dpp)\}_2RhX_2]^{5+}$ (TL = bpy or phen, X = Cl or Br) systems possess ^{3}MLCT emissive excited states [54–56]. Model systems are used for comparison of the excited state properties to better understand the nature of energy and/or electron transfer within the supramolecular architecture. The bimetallic complex $[\{(bpy)2Ru\}_2(dpp)]^{4+}$ that lacks the Rh^{III} typically serves as the model for the interpretation of the emissive properties of these supramolecules due to the similar excited state energies and nature of the emissive ^{3}MLCT excited states. Determination of the rate of electron transfer requires that the energy and nature of the emissive state of the model must closely match that of the supramolecular assembly. In the $[\{(TL)_2Ru(dpp)\}_2RhX_2]^{5+}$ (TL = bpy, X = Cl or Br) and $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$ systems, the optically excited ^{1}MLCT state populates a ^{3}MLCT state that is quenched by intramolecular electron transfer to populate a low-lying ^{3}MMCT state (see Fig. 11) [54–56]. The complexes $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$, $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$, and $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$ display weak emissions from the Ru(dπ) → dpp(π*) ^{3}MLCT state at $\lambda_{max}{}^{em}$ = 760 nm for $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$ and $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$ and $\lambda_{max}{}^{em}$ = 746 nm for $[\{(phen)_2Ru$



Hydrogen via Direct Solar Production. Figure 11
State diagrams for $[\{(bpy)_2Ru(dpp)\}_2RhX_2]^{5+}$ (X = Cl^- or Br^-) and $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$ (*left*) and $[\{(bpy)_2Ru\}_2(dpp)]^{4+}$ (*right*)

(dpp)}$_2$RhCl$_2$]$^{5+}$ with Φ^{em} = 7.3 × 10^{-5}, 1.5 × 10^{-4}, and 1.8 × 10^{-4}, respectively, at RT in deoxygenated acetonitrile solutions following excitation at 520 nm. The emission from the ^{3}MLCT state is quenched by *ca.* 85–90% relative to [{(bpy)$_2$Ru}$_2$(dpp)]$^{4+}$ ($\lambda_{max}{}^{em}$ = 744 nm, Φ^{em} = 1.38 × 10^{-3}) indicating intramolecular electron transfer to populate a lower lying state. Electrochemical analysis predicts this state to be a ^{3}MMCT excited state with oxidized Ru LAs and a reduced Rh. The τ of the ^{3}MLCT states in deoxygenated acetonitrile solutions at RT for [{(bpy)$_2$Ru(dpp)}$_2$RhCl$_2$]$^{5+}$, [{(bpy)$_2$Ru(dpp)}$_2$RhBr$_2$]$^{5+}$, and [{(phen)$_2$Ru(dpp)}$_2$RhCl$_2$]$^{5+}$ are 32, 26, and 27 ns, respectively, reduced relative to the τ of [{(bpy)$_2$Ru}$_2$(dpp)]$^{4+}$, τ = 140 ns [56, 84]. The k_{et} to populate the ^{3}MMCT states, estimated by assuming k_r and k_{nr} of the trimetallic systems are the same as [{(bpy)$_2$Ru}$_2$(dpp)]$^{4+}$, are comparable with k_{et} = 1.2 × 10^8 s^{-1}, 5.2 × 10^7 s^{-1}, and 4.4 × 10^7 s^{-1}, respectively, for [{(bpy)$_2$Ru(dpp)}$_2$RhCl$_2$]$^{5+}$, [{(bpy)$_2$Ru(dpp)}$_2$RhBr$_2$]$^{5+}$, and [{(phen)$_2$Ru(dpp)}$_2$RhCl$_2$]$^{5+}$. Efficient intramolecular electron transfer is seen in this trimetallic motif.

Rhodium-Based Supramolecular Photocatalysts: Solar Hydrogen Production Solar hydrogen production from water using a series of supramolecular photoinitiated electron collectors of the general form LA-BL-RhX$_2$-BL-LA has been evaluated [54–56]. These systems photoreduce by two electrons, following visible light excitation in the presence of DMA. This mechanism proposes that [{(bpy)$_2$Ru(dpp)}$_2$Rh^I]$^{5+}$ species is formed by the excited trimetallic molecule using a sacrificial electron donor and passes through an intermediate Rh^{II} species, Fig. 12. These complexes are unusually oxidizing in their excited states, allowing for excited state oxidation of a variety of electron donors. Reductive quenching of both the ^{3}MLCT and ^{3}MMCT states are feasible. Kinetic investigation by Stern-Volmer emission quenching as well as product


Ru^{II} dpp Rh^{III} dpp Ru^{II}
DMA
DMA
k$_2$
k$_1$
hν
1st excitation
k$_4$
DMA
DMA
k$_3$
*Ru^{III} dpp$^-$ Rh^{III}dpp Ru^{II}
k$_{et}$
*Ru^{III} dpp Rh^{II}dpp Ru^{II}
DMA
k$_q$
DMA$^+$
k$_{q2}$
DMA
DMA$^+$
Ru^{II} dpp Rh^{II} dpp Ru^{II}
DMA
DMA
k$_2$′
k$_1$′
hν
2nd excitation
k$_4$′
k$_3$′
DMA
DMA
*Ru^{III} dpp$^-$ Rh^{II}dpp Ru^{II}
k$_{et}$′
*Ru^{III} dpp Rh^{I}dpp Ru^{II}
DMA
k$_q$′
DMA$^+$
k$_{q2}$′
DMA
DMA$^+$
Ru^{II} dpp Rh^I dpp Ru^{II}
disproportionation
Ru^{II} dpp Rh^{II} dpp Ru^{II}
Ru^{II} dpp Rh^{III} dpp Ru^{II}

Hydrogen via Direct Solar Production. Figure 12
Proposed mechanism for photoinitiated electron transfer. bpy, Cl, and charges are omitted for clarity. Ru^{II}dppRh^{III}dppRu^{II} is [{(bpy)$_2$Ru(dpp)}$_2$RhCl$_2$]$^{5+}$, where bpy = 2,2.-bipyridine and dpp = 2,3-bis-2-pyridylpyrazine. DMA is dimethylaniline. k$_1$ and k$_4$ are the unimolecular decay pathways from the ^{3}MLCT and ^{3}MMCT states, respectively, whereas k$_2$ and k$_3$ are the bimolecular deactivation rates from the same states. k$_{et}$ is the rate of electron transfer to generate the ^{3}MMCT state. k$_q$ and k$_{q2}$ are the rates of reductive quenching by DMA to generate the Rh^{II} complex from the ^{3}MLCT and ^{3}MMCT states, respectively

formation was performed focusing on the formation of the single electron reduced Rh^{II} species. The rate of reductive quenching to generate the Rh^{II} species from the ^{3}MLCT excited state would be $1.9 \times 10^9\ M^{-1}\ s^{-1}$ if reduction only occurs via this pathway [82]. A Stern-Volmer emission quenching analysis showed that the ^{3}MLCT excited state is efficiently quenched by DMA close to diffusion limit, $k_q = 2 \times 10^{10}\ M^{-1}\ s^{-1}$. The estimated excited state reduction potentials of the ^{3}MLCT and ^{3}MMCT states for $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$ are 1.23 and 0.84 V vs. SCE, respectively. Positive driving forces for reductive quenching of both excited states are predicted, using Eq. 10, the oxidation potential of DMA, $E_{1/2} = 0.81$ V vs. SCE. Very similar driving forces for reductive quenching of the excited states by DMA are also predicted for $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$ and $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$ based on their excited state reduction potentials and the oxidation potential of DMA.

The trimetallic complexes $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$, $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$, $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$, $[\{(bpy)_2Os(dpp)\}_2RhCl_2]^{5+}$, $[\{(tpy)RuCl(dpp)\}_2RhCl_2]^{3+}$, $[\{(tpy)OsCl(dpp)\}_2RhCl_2]^{3+}$, and the first photoinitiated electron collector $[\{(bpy)_2Ru(dpb)\}_2IrCl_2]^{5+}$ were evaluated with respect to their functioning as hydrogen photocatalysts in the presence of water and DMA, Et_3N, or TEOA [56]. All complexes demonstrate photocatalytic activity with the exceptions of $[\{(bpy)_2Ru(dpb)\}_2IrCl_2]^{5+}$ and $[\{(tpy)OsCl(dpp)\}_2RhCl_2]^{3+}$. The lack of photocatalysis by $[\{(bpy)_2Ru(dpb)\}_2IrCl_2]^{5+}$ demonstrate that the functioning systems require a Rh center and a ^{3}MLCT and/or ^{3}MMCT states with sufficient driving forces for excited state reduction by the ED. It is predicted that the coordinatively unsaturated species generated by Rh reduction, $[\{(TL)_2Ru(dpp)\}_2Rh^I]^{5+}$ (TL = bpy or phen), is reactive and susceptible to interaction with substrates such as water.

Photolysis of acetonitrile solutions of $[\{(bpy)_2Ru(dpp)\}_2RhX_2]^{5+}$ ($X = Cl^-$ or Br^-) or $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$ in the presence of DMA and water at $\lambda = 470$ nm leads to water reduction to produce hydrogen with $\Phi = 0.01$ with 18–26 TON in 2 h (36–54 TON relative to the LA as 2 TON of the LA are needed per catalytic cycle). $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$ shows higher photocatalytic activity relative to the chloride analog implying that the rate of halide loss is kinetically important. The complexes $[\{(bpy)_2Os(dpp)\}_2RhCl_2]^{5+}$ and $[\{(tpy)RuCl(dpp)\}_2RhCl_2]^{3+}$ yield similar amounts of hydrogen when irradiated at 470 nm in the presence of DMA and water, but with much lower efficiency than $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$. The excited state reduction potentials of the ^{3}MLCT and ^{3}MMCT were predicted as 0.91 and 0.54 V vs. SCE, respectively, for $[\{(bpy)_2Os(dpp)\}_2RhCl_2]^{5+}$ and 1.01 and 0.61 V, respectively, for $[\{(tpy)RuCl(dpp)\}_2RhCl_2]^{3+}$ making quenching of the ^{3}MMCT excited state by DMA unlikely.

The production of hydrogen despite the negative driving force for reductive quenching of the ^{3}MMCT excited state suggests that hydrogen production can occur through quenching of only the ^{3}MLCT excited state. The lower driving force for reductive quenching of the ^{3}MLCT and the lower τ may account for the lower photocatalytic efficiency of these complexes. $[\{(tpy)OsCl(dpp)\}_2RhCl_2]^{3+}$ does not function as a photocatalyst under the conditions investigated. Based on the excited state reduction potentials of the ^{3}MLCT and ^{3}MMCT states, 0.71 and 0.37 V vs. SCE, respectively, the reductive quenching of both excited states by DMA are thermodynamically unfavorable, explaining the lack of photocatalytic activity of this system. Addition of Hg(l) to the photocatalytic systems does not impair hydrogen production suggesting that hydrogen production catalyzed by Rh colloid formation is not an active pathway and that the supramolecular architecture remains intact during photocatalysis [54, 86, 87]. Interestingly, $[\{(bpy)_2Ru(dpb)\}_2IrCl_2]^{5+}$ is not a solar hydrogen photocatalyst although PEC occurs on the dpb(π^*) orbitals and a thermodynamically favorable driving force for reductive quenching of the ^{3}MLCT state by DMA exists. This implies that other factors in addition to PEC and favorable thermodynamics for reductive quenching are important for solar hydrogen production. The lack of hydrogen production by the Ir-system implies that the Rh center is essential for photocatalysis.

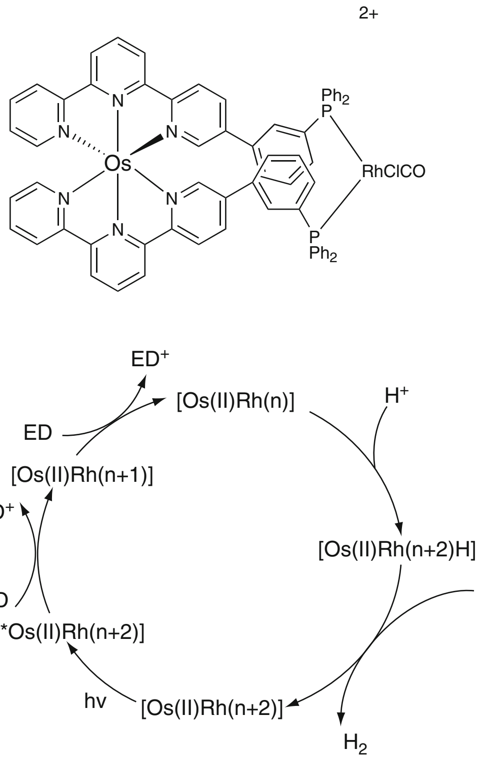
The impact of the ED on the most efficient photocatalysts $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$, $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$, and $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$ shows that the efficiency of hydrogen production varies in the order DMA (18–26 TON in 2 h) > Et_3N (3–6 TON in 2 h) > TEOA (1 TON in 2 h) [56]. TEOA has a slightly higher driving force for reductive quenching of the photocatalysts than Et_3N, but results in lower hydrogen yields. An important factor to be considered in water reduction is the solution pH. The effective pH of the photolysis solutions using DMA, Et_3N, or TEOA were estimated as *ca.* 9.1, 14.7, and 11.8, respectively, on the assumption that the pK_a values of their conjugated acids remain unchanged in the photocatalytic solutions relative to aqueous conditions (pK_a = 5.07 ($DMAH^+$), 10.75 (Et_3NH^+), and 7.76 ($TEOAH^+$)). The lower catalyst function when TEOA is the ED may be attributed to the higher effective pH of the solution. The reduction potential of water is pH dependent occurring at lower potentials at lower pH. In addition the protonation of the Rh center may be critical to photocatalysis and this equilibrium would be very sensitive to pH. In addition, the ability of DMA to form donor-LA π-stacking interactions may be advantageous for more efficient reductive quenching of the excited states affording the highest hydrogen yields compared to the aliphatic electron donors. Recent studies have focused on the scale up of photochemical hydrogen production using the lead photocatalyst $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$. Studies show that increased [DMA], increased headspace, volume, and a DMF solvent provide enhanced catalyst function [88].

Photochemical properties of $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]Br_5$ in aqueous medium have been investigated [57]. Studies have established this complex to function as a photoinitiated electron collector in water, the first system shown to display this function in aqueous medium. $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]Br_5$ functions as a photocatalyst for the production of hydrogen in the presence of TEOA with added triflic acid, hydrobromic acid, or phosphoric acid. The photocatalytic efficiency is lower in the aqueous medium, typical of these systems.

The Rh-centered supramolecular complexes are efficient photocatalysts for water reduction to produce hydrogen fuel. The variation of component identity within this molecular architecture allows for optimization of functioning of these systems and the understanding of the mechanism of action. The functioning of these systems is also modulated by system parameters such as solvent, temperature, reagent concentrations, hydrogen partial pressure, etc. This provides a unique forum to explore the multidimensional space needed to optimize solar hydrogen production via direct water reduction. Often preliminary studies appear without the long-term study needed to explore the full potential of these new supramolecular catalysts for hydrogen production.

Supramolecular solar hydrogen photocatalysts that couple a $[Os(tpy)_2]^{2+}$ LA to a Rh CAT are reported, Fig. 13 [58]. The OsRh dyads have been synthesized and analyzed in situ by treatment with $[RhCl(CO)_2]_2$ or $RhCl_3$. The dyads are potent LAs absorbing in the visible at 480 and 655 nm, attributed to Os(dπ) → tpy (dπ*) ^{1}MLCT and ^{3}MLCT absorptions, respectively. An Os(dπ) → tpy(dπ*) ^{3}CT emission is observed at 728 nm (Φ^{em} = 0.52 in an acetonitrile/water mixture) with reduced intensity relative to the parent LA (Φ^{em} = 0.73 in an acetonitrile/water mixture) suggestive of intramolecular electron transfer to the Rh. The emission is decreased when sodium ascorbate is added to the dyad efficient quenching of the excited state by sodium ascorbate.

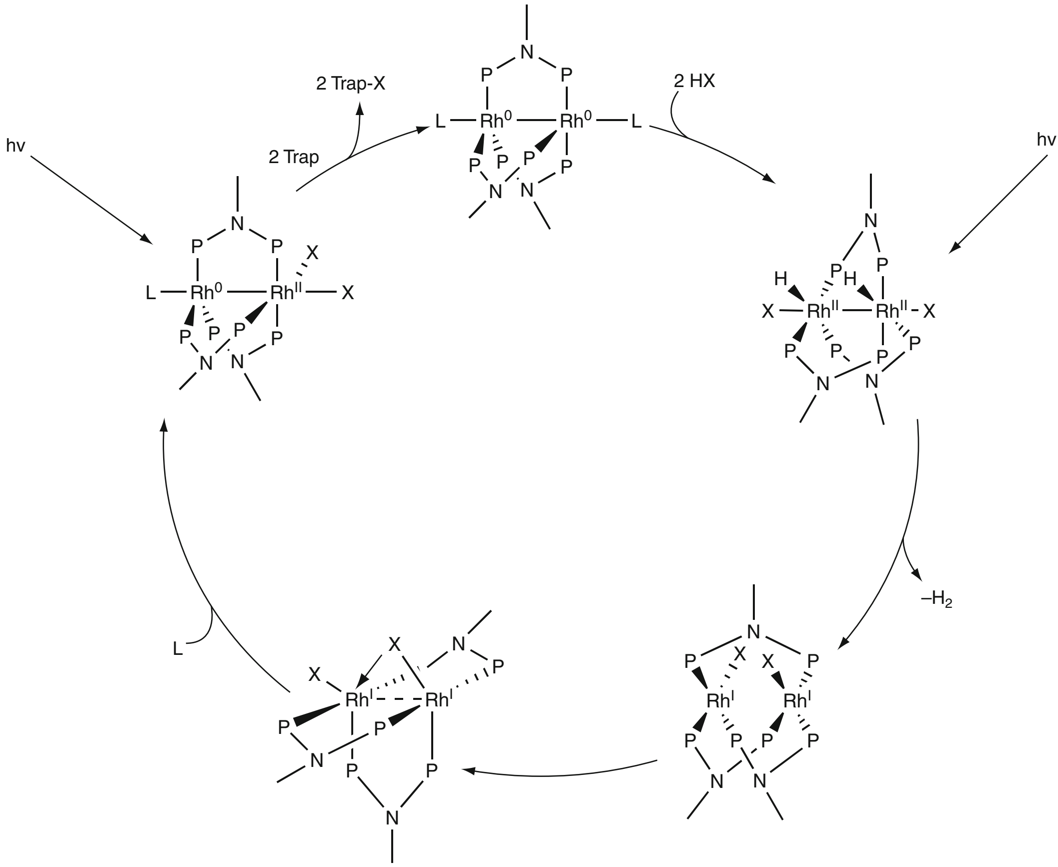
Irradiation of catalytic amounts of the dyad generated in situ at λ > 380 nm in a trifluoromethanesulfonic



Hydrogen via Direct Solar Production. Figure 13
OsRh dyad systems for solar hydrogen production

acid/sodium ascorbate mixture at pH = 5.2 for 18 h affords hydrogen with 36 TON [58]. Use of $RhCl_3$ in place of $[RhCl(CO)_2]_2$ yields 87 TON with $\Phi = 0.007$. Electron donors including TEOA or L-cysteine are not effective for hydrogen photocatalysis. Lower catalytic function is observed when $RhCl_3{\bullet}3H_2O$/dppe (dppe = 1,2-bis(diphenylphosphino)ethane) with $[Os(tpy)_2]^{2+}$ is used as the photosystem (24 TON) with no detectable hydrogen if only $RhCl_3{\bullet}3H_2O$/dppe, RhCl(CO)(dppe), or only the LA is used. This suggests that the supramolecular architecture is important for photocatalysis and intramolecular electron transfer from Os to Rh is a key step. A schematic for the proposed mechanism is shown in Fig. 13. It is proposed that upon optical excitation, intramolecular electron transfer to Rh by two electrons leads to a reduced Rh which can react with a proton to form a Rh-H, which generates hydrogen.

Photochemical hydrogen production from hydrohalic acid solutions by a Rh-based dinuclear mixed-valence complex capable of multielectron photochemistry has been reported, Fig. 14 [59, 89]. Hydrogen production occurs with $\Phi = 0.01$ with 27 TON/h during the initial 3 h, in the presence of a halogen trap. The mixed-valence compounds have been constructed to drive multielectron chemistry and are of the form $M^{n+}\cdots M^{n+2}$. Rhodium bimetallics, $[Rh_2(dfpma)_3X_4]$, when ligated by three dfpma ligands transform to a two-electron mixed-valence state, $[Rh_2^{0,II}(dfpma)_3X_2(L)]$ (dfpma = $MeN(PF_2)_2$, X = Cl or Br, L = CO,



Hydrogen via Direct Solar Production. Figure 14
Mechanism for the photocatalytic generation of hydrogen from hydrohalic acids using a dirhodium mixed-valence photocatalyst (Figure adapted from Ref. [59])

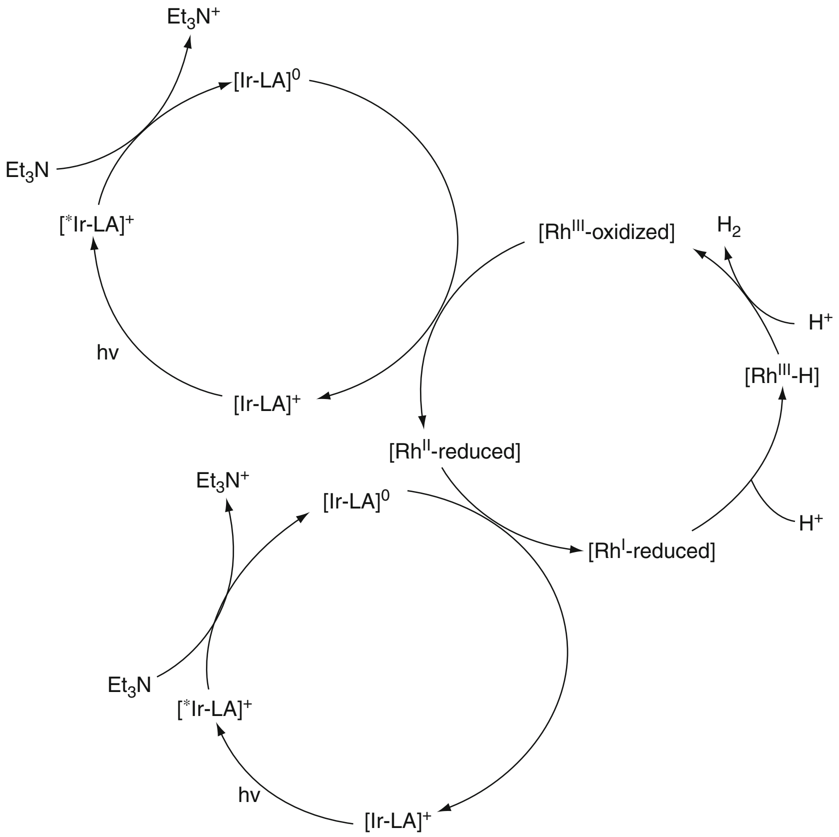
PR_3, or CNR) when irradiated at excitation wavelengths between 300 and 400 nm in the presence of excess L and a halogen-atom trap. The dfpma ligand provides stability to the mixed-valence state by acting as both a π-acceptor and a π-donor. Further irradiation activates a Rh^{II}-X, which is the rate limiting step, to generate a doubly reduced form, $[Rh_2^{0,0}(dfpma)_3 L_2]$. Photolysis of $[Rh_2^{0,0}(dfpma)_3 L_2]$ in the presence of HCl results in an intermediate Rh^{II},Rh^{II} dihydride, dihalide, $[Rh_2^{II,II}(dfpma)_3Cl_2H_2]$, which upon photolysis produces hydrogen with the generation of the $[Rh_2^{I,I}(dfpma)_3Cl_2]$ which undergoes internal disproportionation affording $[Rh_2^{0,II}(dfpma)_3X_2(L)]$. New systems $[Rh^IAu^I(tfepma)_2(CN^tBu)_2]^{2+}$ and $[Pt^{II}Au^I(dppm)_2PhCl]^+$ (tfepma = $MeN(P(OCH_2CF_3)_2)_2$ and dppm = $CH_2(PPh_2)_2$) have been constructed with the aim of promoting more efficient M-X bond activation [90, 91]. The Pt^{III}-Au^{II} species, $[Pt^{III}Au^{II}(dppm)_2PhCl_3]^+$, formed by photooxidation displays enhanced efficiency of metal-halide bond activation with respect to the dirhodium complexes.

Rhodium-Based Multicomponent Photocatalysis The pioneering work in photochemical hydrogen production using molecular systems used $[Rh(bpy)_3]^{3+}$ as an electron relay in solar hydrogen production schemes using $[Ru(bpy)_3]^{2+}$ and TEOA [11, 60]. In the presence of a heterogeneous platinum catalyst, photocatalytic hydrogen production from water occurs with $\Phi = 0.11$ irradiated at $\lambda = 450$ nm with *ca.* 80 TON relative to $[Rh(bpy)_3]^{2+}$ reported under similar conditions [11, 60]. The MLCT excited state of $[Ru(bpy)_3]^{2+}$ undergoes intermolecular electron transfer to $[Rh(bpy)_3]^{3+}$ to produce the oxidized LA and $[Rh(bpy)_3]^{2+}$. The photogenerated Rh^{II} system is kinetically unstable and can rapidly disproportionate to $[Rh^I(bpy)_2]^+$ and $[Rh^{III}(bpy)_3]^{3+}$ deactivating the system. The catalytic species in this system is colloidal Pt, which facilitates hydrogen photocatalysis by accepting electrons from the reduced Rh. In-depth studies of $[Rh^I(bpy)_2]^+$ have established that at low concentrations in alkaline solutions $[Rh^I(bpy)_2]^+$ predominates, while a dimer, $[Rh^I(bpy)_2]_2^{2+}$, is dominant at higher $[Rh^I(bpy)_2]^+$ [92, 93]. $[Rh^{III}(bpy)_2(H)(H_2O)]^{2+}$ and $[Rh^I(bpy)_2]_2(H)^{3+}$ are formed by protonation of $[Rh^I(bpy)_2]^+$ or $[Rh^I(bpy)_2]_2^{2+}$ at low pH. Thermodynamic and kinetic parameters for the oxidative addition of hydrogen to $[Rh^I(bpy)_2]^+$ and photoinduced reductive elimination of hydrogen have been studied, $K_H = 1.45 \times 10^3\ M^{-1}$ [94]. Studies show that *cis*-$[Rh^{III}(H)_2(bpy)_2]^+$ dissociates to hydrogen and $[Rh^I(bpy)_2]^+$ when irradiated ($\lambda > 300$ nm).

Cyclometalated Ir^{III} complexes have been used as LAs in photochemical schemes for production of hydrogen from water using tris-2,2′-bipyridyl Rh^{III} complexes as CATs, Fig. 15 [61]. The study of the system function with varied $[Ir(C\wedge N)_2(N\wedge N)]^+$ LAs and $[Rh(N\wedge N)]^{3+}$ CATs has allowed the development of an optimized photosystem for solar hydrogen production. An optimal catalyst combination of $[Ir(f\text{-}mppy)_2(dtbbpy)]^+$ and $[Rh(dtbbpy)_3]^{3+}$ has been identified through catalyst screening, studies of the impact of hydrogen production by LA, CAT, solvent environment, and electron donor (f-mpy = 5-methyl-2-(4-fluoro)phenylpyridine). Irradiation of 80% THF-H_2O system at 460 nm shows 5,000 TON relative to either the LA or the CAT when one component is limiting in excess of the other. The use of 1:1 of LA and CAT balances the performance of both components resulting in 3,400 TON and a maximum catalytic rate of turnovers of 520 h^{-1} with $\Phi = 0.32$ in *ca.* 20–24 h. The rate constants for reductive quenching of the excited state of the LA by Et_3N and oxidative quenching by $[Rh(dtbbpy)_3]^{3+}$ are similar, $k_{ox} = 6.7 \times 10^7\ M^{-1}\ s^{-1}$ and $k_{red} = 6.4 \times 10^7\ M^{-1}\ s^{-1}$, but the presence of excess of Et_3N suggests that a reductive quenching mechanism is dominant. Addition of Hg does not impede system function providing evidence that the system is a true homogeneous system for photocatalytic hydrogen production.

Co-Based Supramolecular and Multicomponent Systems

Cobalt complexes with diglyoxime ligands are electrocatalysts for hydrogen production at low overpotentials. The electrochemical and photochemical hydrogen evolution catalyzed by cobaloximes with thermodynamic analysis of hydrogen evolution pathways has been reviewed recently [95]. Supramolecular systems as well as multicomponent systems which incorporate cobaloximes sensitized by Ru or Ir base LAs are known. These systems provide a means to utilize first-row transition metal complexes as CATs.

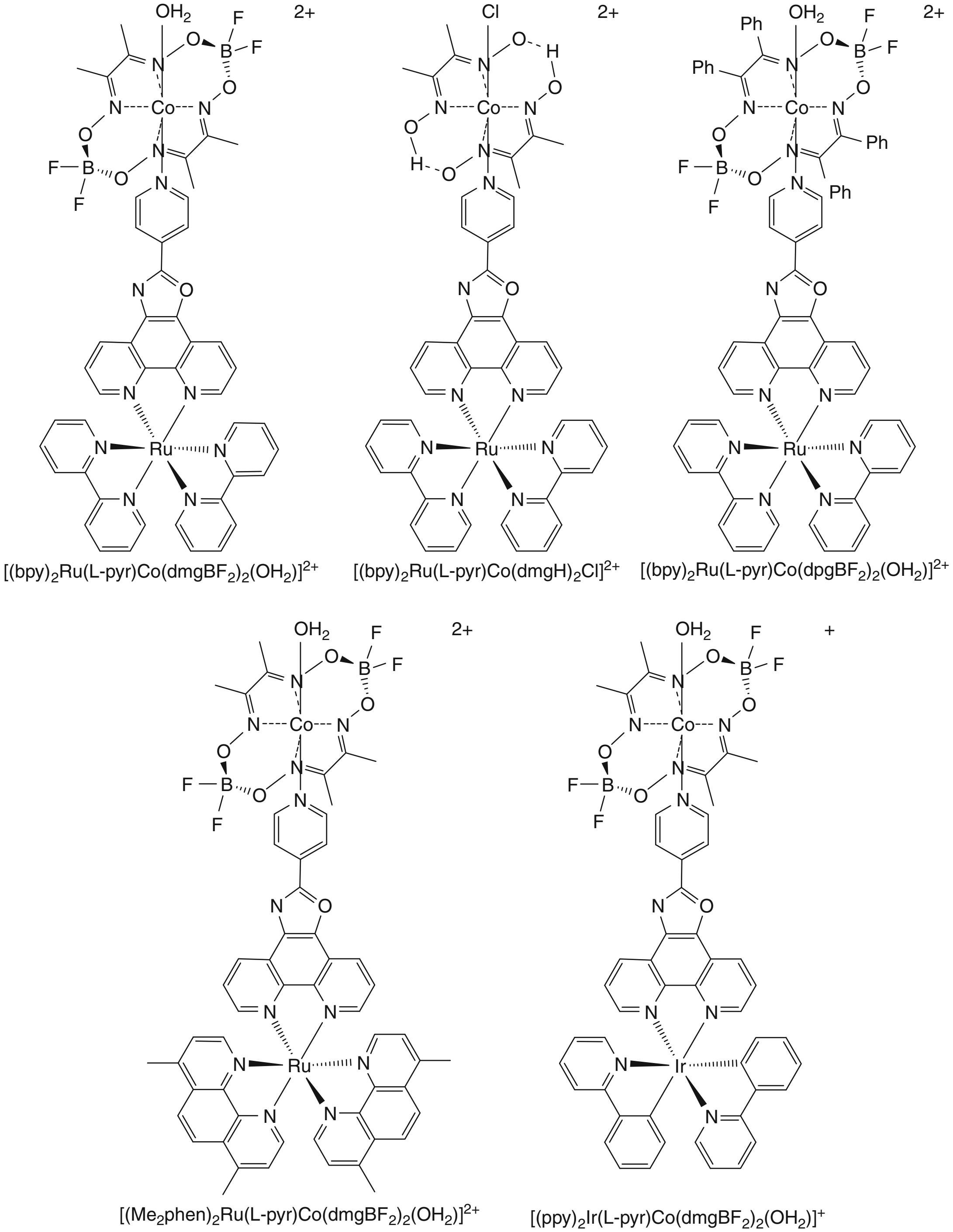


Hydrogen via Direct Solar Production. Figure 15

Mechanism for the photocatalytic generation of hydrogen from water in a Ir-Rh multicomponent system with Et_3N electron donor

Co-Based Supramolecular Systems Supramolecular systems have been engineered taking advantage of the light-absorbing properties of Ru^{II} or Ir^{III} chromophores and the catalytic properties of cobaloximes in single molecules [62, 63]. Supramolecular dyads $[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$, $[(bpy)_2Ru(L\text{-}pyr)Co(dmgH)_2Cl]^{2+}$, $[(bpy)_2Ru(L\text{-}pyr)Co(dpgBF_2)_2(OH_2)]^{2+}$ [62], $[(Me_2phen)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$, and $[(ppy)_2Ir(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{+}$ (L-pyr = [(4-pyridine)oxazolo(4,5-f)phenanthroline], dmgH = dimethylglyoximate, $dmgBF_2$ = (difluoroboryl)dimethylglyoximate, $dpgBF_2$ = (difluoroboryl)diphenylglyoximate, Me_2phen = 4,7-dimethylphenanthroline, ppy = 2-phenylpyridine) [63] are photocatalysts for solar hydrogen production, Fig. 16. The BF_2-bridged systems are more resistant toward acid hydrolysis and undesired hydrogenation reactions. The Co^{II} state is easier to reduce in the BF_2-bridged systems vs. the H-bridged cobaloximes. The L-pyr BL facilitates intramolecular excited state electron transfer.

Photochemical hydrogen production from Et_3NH^+ occurs when acetone solutions of $[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$, $[(bpy)_2Ru(L\text{-}pyr)Co(dmgH)_2Cl]^{2+}$, or $[(bpy)_2Ru(L\text{-}pyr)Co(dpgBF_2)_2(OH_2)]^{2+}$ are photolyzed using a CdI-doped Hg light source in the presence of 100-fold excesses of Et_3N and Et_3NH^+ [62]. Turnover numbers of 56, 17, and 12 over 4 h are seen for $[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$, $[(bpy)_2Ru(L\text{-}pyr)Co(dmgH)_2Cl]^{2+}$, or $[(bpy)_2Ru(L\text{-}pyr)Co(dpgBF_2)_2(OH_2)]^{2+}$, respectively. Increasing Et_3N and Et_3NH^+ to 300-fold excess had slight effect with 60 TON observed for $[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$ in 4 h. The impact of solvent on photocatalytic efficiency of $[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$ showed lower TON in acetonitrile



$[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$ $[(bpy)_2Ru(L\text{-}pyr)Co(dmgH)_2Cl]^{2+}$ $[(bpy)_2Ru(L\text{-}pyr)Co(dpgBF_2)_2(OH_2)]^{2+}$

$[(Me_2phen)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$ $[(ppy)_2Ir(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{+}$

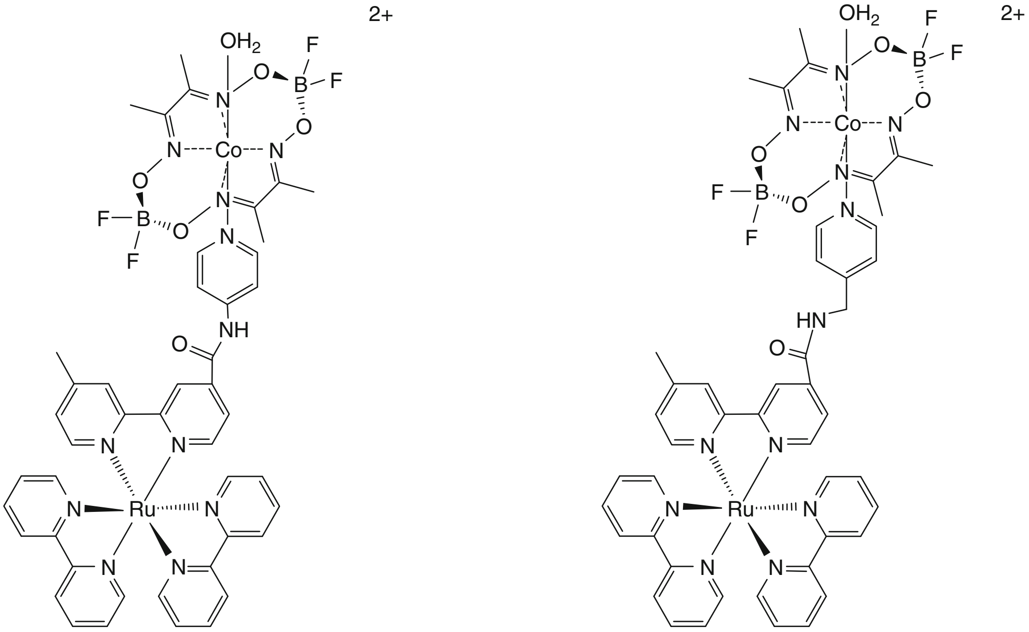
Hydrogen via Direct Solar Production. Figure 16
Representations of cobalt-based supramolecular dyads

(TON = 10), methanol (TON = 9), DMF (TON = 3), or 1,2-dichloroethane (TON = 0). The lower TON in DMF or 1,2-dichloroethane is attributed to decomplexation of the photocatalyst in these solvents. When the proton source is water (100-fold excess of both water and Et_3N), 22 TON in 4 h is seen for $[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$. It is likely that the lower activity is due to pH changes that are significant when water is used as the proton source. Visible light-induced excitation ($\lambda > 350$ nm) of the $[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$ system afforded 103 TON in 15 h. Lower TON is obtained when a UV cutoff filter is used, implying both UV and visible light promote photocatalysis. It is mentioned that the analogous system of $[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$ with the BL oxazole O replaced by NH shows a twofold increased photocatalytic activity with 104 TON achieved in 4 h using 100-fold excesses of Et_3N and Et_3NH^+. For comparison, photolysis of the multicomponent system consisting of $[Ru(bpy)_3]Cl_2$ and one equivalent of $[Co(dmgBF_2)_2(OH_2)_2]$ in 100-fold excess of both Et_3N and Et_3NH^+ yields only 20 TON using UV-VIS light with only traces of hydrogen observed when a UV cutoff filter is used.

Photochemical hydrogen production occurs when acetone solutions of $[(Me_2phen)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$ (Me_2phen = 4,7-dimethylphenanthroline), in which bpy has been substituted by Me_2phen, or $[(ppy)_2Ir(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^+$, in which an Ir-based LA is used, are irradiated at $\lambda > 380$ nm in the presence of 300 equivalents of Et_3N and Et_3NH^+ with 9 and 90 TON in 4 h, respectively [63]. The Ir-based system demonstrates enhanced photocatalytic efficiency to the Ru-based system, with 210 turnovers in 15 h in the presence of 600 equivalents of ED and proton source. For comparison the multicomponent system $[Ir(ppy)_2(phen)]^+$ and $[Co(dmgBF_2)_2(OH_2)_2]$ under the same irradiation conditions yields similar initial rates, but the dyad shows greater stability as the hydrogen production rate levels off at 165 TON in 15 h. This implies that intramolecular electron transfer in the dyad may not be the rate limiting step in the hydrogen production scheme, but the reductive quenching of the excited photosensitizer by Et_3N may be important.

This study also uses a multicomponent system based on a $[Re(CO)_3Br(phen)]$ LA and $[Co(dmgBF_2)_2(OH_2)_2]$ CAT with 273 TON in 15 h using 600-fold excess of both Et_3N and Et_3NH^+ [63]. The use of the Re-based LA provides enhanced system functioning with Φ (½ H_2) = 0.16 relative to Φ (½ H_2) = 0.10 in the Ir-based multicomponent system (λ = 412 nm, irradiation for 2 h). This study demonstrates superior system functioning when Ir- or Re-based LAs are employed relative to the Ru-based LAs. The Ru-based dyads demonstrate superior photocatalytic efficiency to the multicomponent $[Ru(bpy)_3]^{2+}$/cobaloxime system. By contrast, the Ir-based dyad shows similar catalytic efficiency to the multicomponent $[Ir(ppy)_2(phen)]^+$/cobaloxime system, but the supramolecular dyad shows greater stability. It is predicted that intramolecular electron transfer affords a Co^I species which is protonated to yield a Co^{III}-H intermediate that can be further protonated to generate hydrogen. Electron transfer to the catalytic Co center can occur either directly from oxidative quenching of the photoexcited LA or from a reduced LA generated from the reductive quenching by an electron donor. A favorable driving force for oxidative quenching of the LA by the attached Co^{II} site is predicted, as the observed electrocatalytic potentials are more positive than the standard excited state reduction potential for $[Ru(bpy)_3]^{2+}$ [96]. Thus, the mechanism for photochemical hydrogen production by the supramolecular Ru-based systems is predicted to proceed through intramolecular electron transfer to the cobalt center from a photoexcited Ru LA. Faster reductive quenching of the excited Re- or Ir-based LA by Et_3N is attributed to the enhanced hydrogen production observed when the photosystem uses these LAs.

The complex $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$ in which the Ru LA and the cobaloxime CAT are connected by a conjugated bridge, and $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH(CH_2)\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$, in which the two components are connected by an unconjugated bridge, have been used in solar hydrogen production schemes, Fig. 17 [64]. Intramolecular electron transfer is evidenced by the emission quenching of the ^{3}MLCT charge-transfer state by 35% and 25% relative to the parent LAs $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH\text{-}4\text{-}py)]^{2+}$ and $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONHCH_2\text{-}4\text{-}py)]^{2+}$,



$[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$

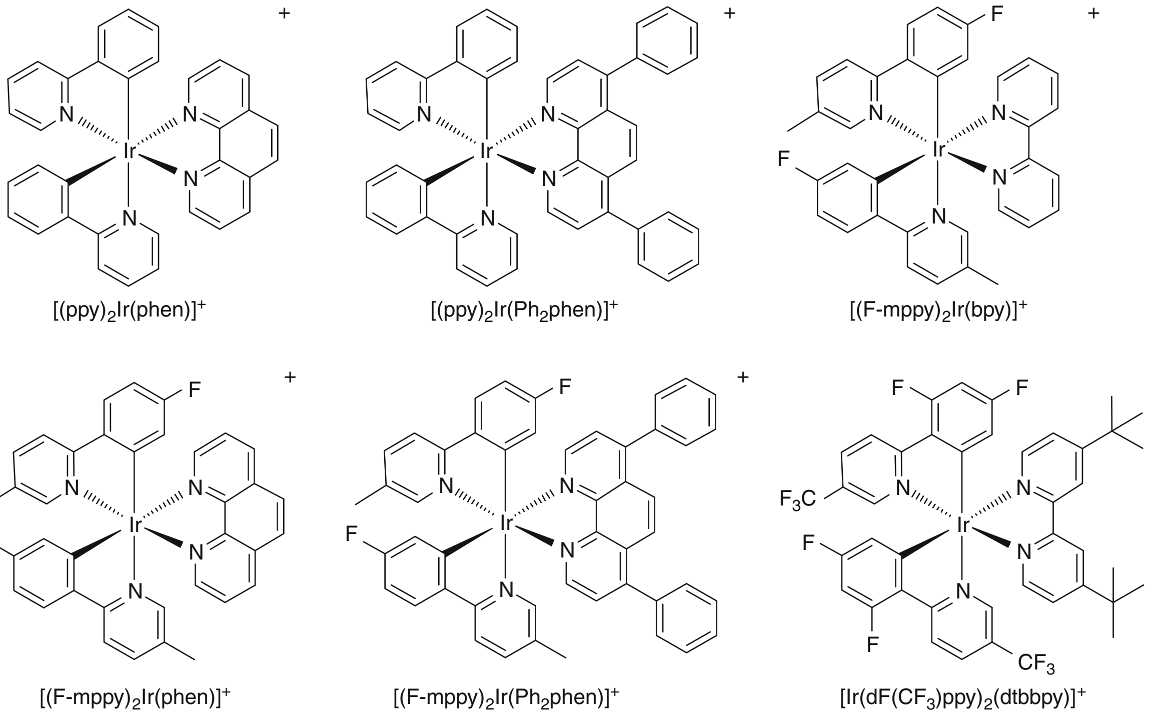
$[(bpy_2)Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH(CH_2)\text{-}4\text{-}py)Co(dmgBF_2)_2(OH2)]^{2+}$

Hydrogen via Direct Solar Production. Figure 17
Representations of $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$ and $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH(CH_2)\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$

respectively. The lower emission intensity in the dyad containing the conjugated bridge is attributed to more efficient deactivation through radiative and non-radiative decay, due to the shorter distance between the LA and the cobaloxime. Photolysis of $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$ or $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH(CH_2)\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$ at $\lambda = 400$ nm in the presence of 300-fold excess of both Et_3N and Et_3NH^+ in acetone results in hydrogen production with 38 and 48 TON in over 8 h for $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$ and $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH(CH_2)\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$, respectively. Lower TONs are observed when DMF or acetonitrile are the solvent. The complex $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$ with a conjugated bridge shows similar photocatalytic efficiency to $[(bpy)_2Ru(L\text{-}pyr)Co(dmgBF_2)_2(OH_2)]^{2+}$. The unconjugated bridge may impede the possible deactivation routes promoting higher photocatalytic TON. In $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH\text{-}4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$ and $[(bpy)_2Ru(bpy\text{-}4\text{-}CH_3,4'\text{-}CONH(CH_2)4\text{-}py)Co(dmgBF_2)_2(OH_2)]^{2+}$, oxidative quenching of the excited Ru LA by the Co^{II} through intramolecular electron transfer has been suggested as the dominant quenching process based on the favorable driving forces and the fact that the excited states of the parent LAs are not quenched efficiently with Et_3N.

Other Multicomponent Co-Based Systems The pioneering work using Co-based CAT for solar hydrogen production was in the late 1970s utilizing $[Ru(bpy)_3]^{2+}$ LA and a Co(II) macrocycle system as the CAT. Optical excitation of a system containing $[Ru(bpy)_3]^{2+}$, $[Co^{II}L]$ ($[Co^{II}L] = Co^{II}(Me_6\ [14]dieneN_4)(H_2O)_2)^{2+}$, and Eu^{2+} at $\lambda = 450$ nm afforded hydrogen providing 10 TON with respect to $[Ru(bpy)_3]^{2+}$ and 1 TON with respect to $[Co^{II}L]$, with a $\Phi = 0.05$ [65]. Optical excitation of the LA leads to intermolecular electron transfer generating a Co^I species which reduces water to produce hydrogen. Early work using cobaloximes for photochemical hydrogen production entailed a photosystem consisting of $[Ru(bpy)_3]^{2+}$, $[Co(dmgH)_2(OH_2)_2]$, and triethanolamine (TEOA) in DMF with 16 TON in 1 h relative to the CAT [97].

A series of heteroleptic Ir complexes of the molecular architecture $[Ir(C\text{^}N)_2(N\text{^}N)]^+$ (C^N = ppy or F-mppy and N^N = bpy, phen, Ph_2phen (4,7-diphenyl-1,10-phenanthroline)) have been investigated as LAs in photochemical hydrogen production from water using $[Co(bpy)_3]^{2+}$ as the CAT, Fig. 18 [66]. Irradiation of photosystems consisting of an Ir-based LA, $[Co(bpy)_3]^{2+}$ and TEOA in a CH_3CN/water mixture at $\lambda = 465$ nm affords hydrogen with 800–920 TON relative to the LA and 16–18 TON relative to the CAT with the system consisting of $[Ir(F\text{-}mppy)_2(bpy)]^+$ LA providing the highest TON. The fluorinated compounds typically show slightly higher activity. It is proposed that intramolecular electron transfer from the excited LA to the CAT affords a reduced Co species that promotes proton reduction to make hydrogen. The Ir-based systems are about 20 times more efficient than $[Ru(bpy)_3]^{2+}$ and 3–6 times more efficient than $[Ru(Me_2phen)_3]^{2+}$. The multicomponent Ru-based photosystem consisting of $[Ru(bpy)_3]^{2+}$ or $[Ru(Me_2phen)_3]^{2+}$ (Me_2phen = 4,7-dimethyl-1,10-phenanthroline) provides lower TON (100 TON relative to $[Ru(bpy)_3]^{2+}$ and 2 TON relative to the CAT or 290 TON relative to $[Ru(Me_2phen)_3]^{2+}$ and 12 TON relative to the CAT). The excited states of the Ir-based systems are quenched efficiently by both $[Co(bpy)_3]^{2+}$ and TEOA whereas the excited states of the Ru LAs are only quenched by $[Co(bpy)_3]^{2+}$. The rate of quenching of the excited state of the Ir-system by $[Co(bpy)_3]^{2+}$ is about an order of a magnitude faster than TEOA quenching rate and also faster than for the Ru-based systems. The rates of quenching of the Ir-based excited states by $[Co(bpy)_3]^{2+}$ range from 4.8 to 6.9 × $10^8\ M^{-1}\ s^{-1}$. System modification of the LA has provided $[Ir(dF(CF_3)ppy)_2(dtbbpy)](PF_6)$ (dF(CF_3)ppy) = 2-(2,4-difluorophenyl)-5-trifluoromethylpyridine and dtbbpy = 4,4′-di-tert-butyl-2,2′-dipyridyl) with enhanced excited state lifetime and significant reducing power that shows dramatically improved system function with a relative Φ that is ca. 14 times larger than $[Ru(Me_2phen)_3]^{2+}$ [98].



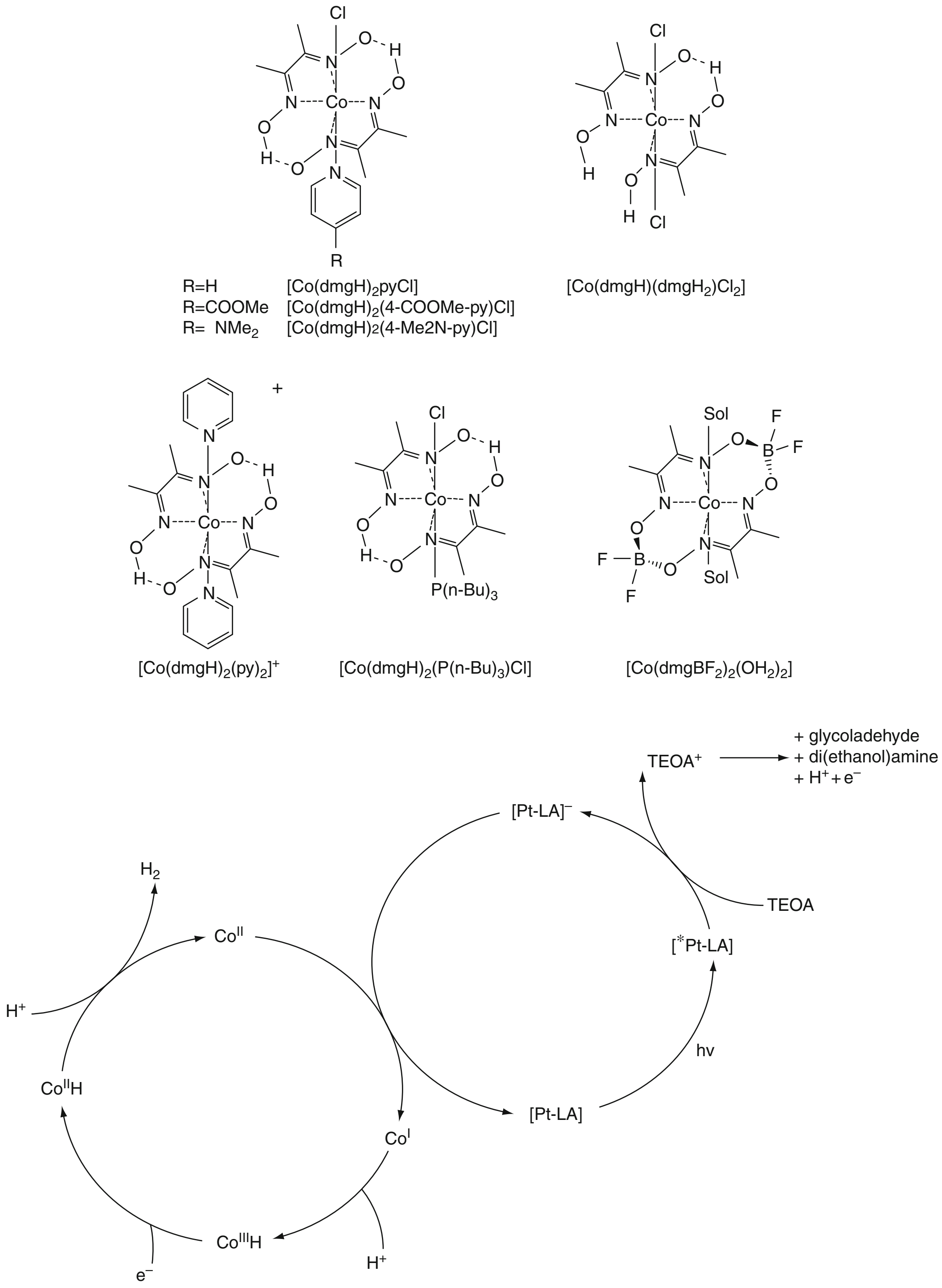
Hydrogen via Direct Solar Production. Figure 18
Ir-based LAs used in photosystems incorporating $[Co(bpy)_3]^{2+}$ CAT for photochemical hydrogen production from water

A series of cobaloxime complexes, [Co(dmgH)$_2$pyCl], [Co(dmgH)$_2$(4-COOMe-py)Cl], [Co(dmgH)$_2$(4-Me$_2$N-py)Cl], [Co(dmgH)(dmgH$_2$)Cl$_2$], [Co(dmgH)$_2$(py)$_2$]$^+$, [Co(dmgH)$_2$(P(n-Bu)$_3$)Cl], and [Co(dmgBF$_2$)$_2$(OH$_2$)$_2$], (dmgH$_2$ = dimethylglyoxime) have been investigated as molecular catalysts photochemical hydrogen generation from water in systems containing a Pt terpyridyl acetylide chromophore and TEOA in acetonitrile, Fig. 19 [67, 99]. Characterization of the cobaloximes, hydrogen photocatalysis, and the investigation of the impact of photosystem functioning with respect to LA, CAT, TEOA, solvent, and CH$_3$CN/water ratio are reported. The cobaloximes oxidatively quench the emission of the excited state LA efficiently, close to the diffusion controlled limit, with the fastest quenching of 1.7×10^9 M^{-1} s^{-1} by [Co(dmgH)$_2$(4-COOMe-py)Cl].

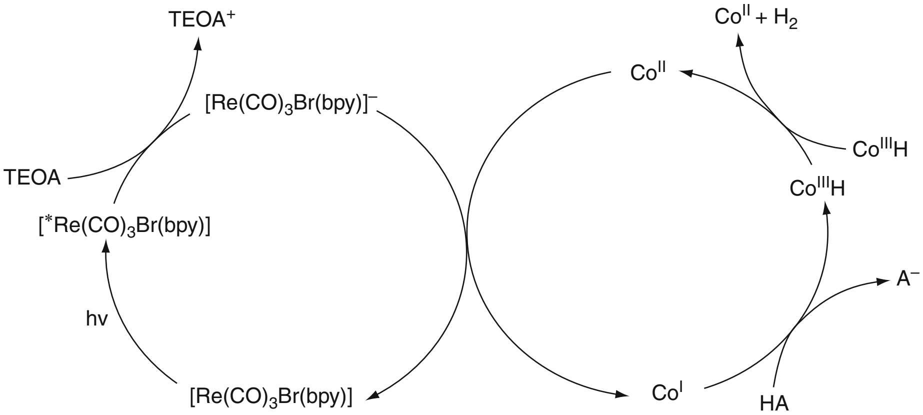
The rate of oxidative quenching by [Co(dmgH)$_2$(py)$_2$]$^+$ is 4.0×10^7 M^{-1}s^{-1}; this lower rate is attributed to the lack of a Cl ligand in this complex that can interact with the Pt LA during the quenching process. It was found that a photosystem consisting of [Pt(ttpy)(C≡CPh)]ClO$_4$ (ttpy = 4′-p-tolyterpyridine), [Co(dmgH)$_2$pyCl], TEOA, and water in CH$_3$CN gives 2,100 TON in 10 h, irradiated at $\lambda > 410$ nm, when the optimized [TEOA] and CH$_3$CN/H$_2$O ratio are used. No hydrogen is observed after 5 h of photolysis in systems consisting of [Pt(ttpy)(C≡CPh)]ClO$_4$, [Co(dmgH)$_2$(P(n-Bu)$_3$)Cl] or [Co(dmgBF$_2$)$_2$(OH$_2$)$_2$], and TEOA. In the system consisting of [Co(dmgH)$_2$(P(n-Bu)$_3$)Cl], it is likely that the stable hydride formed, [CoH(dmgH)$_2$(P(n-Bu)$_3$)] known to yield hydrogen only upon thermolysis at 150°C, prevents hydrogen photocatalysis. The complex [Co(dmgBF$_2$)$_2$(OH$_2$)$_2$] is thermodynamically incapable of proton reduction as evidenced by the Co$^{II/I}$ reduction potential of −0.29 V vs. NHE which is more positive of proton reduction (−0.48 V) at the pH used for photocatalysis (pH = 8.5). Induction periods of 1–2 h are observed for the hydrogen-generating systems. This induction period is impacted by the solvent conditions, LA, and TEOA concentration. The induction period is consistent with the requirement for the formation of Co^I for photogeneration of hydrogen to occur. The photochemical cycle can proceed through the oxidative quenching of the photoexcited LA by Co^{III} to form LA$^+$ and Co^{II}, or through reductive quenching by TEOA to form LA$^-$ and TEOA$^+$, Fig. 19. Electron transfer from the LA$^-$ to Co^{III} occurs regenerating the LA. The catalytic cycle proceeds with further reduction of Co^{II} to Co^I followed by protonation to give a Co^{III}-H intermediate. One pathway is the protonation of Co^{III}-H to yield hydrogen and Co^{III}. An alternative pathway is utilizing a single Co pathway, which entails reduction of the Co^{III}-H to Co^{II}-H which subsequently gets protonated to generate hydrogen and Co^{II}. This single Co mechanism is supported by the observed linear dependence of the Co concentration on photochemical hydrogen production in a system consisting [Pt(ttpy)(C≡CPh)]ClO$_4$ (ttpy = 4′-p-tolyterpyridine), [Co(dmgH)$_2$pyCl], TEOA, and water in CH$_3$CN irradiating at $\lambda > 410$ nm. An alternative kinetically distinguishable proposal is the homolytic reaction mechanism involving two Co^{III}-H to give hydrogen and Co^{II}. Kinetics of electron transfer reactions of hydrogen evolving cobaloximes have been reported recently [100]. Studies demonstrate that the homolytic pathway (Co^{III}H + CoI^{II}H → 2 Co^{II} + H$_2$) is favored over a pathway involving protonation of Co^{III}-H to yield hydrogen and Co^{III}.

A photocatalytic system using the organic dye, Eosin Y as the LA, [Co(dmgH)$_2$pyCl]$^{2+}$, and TEOA produces hydrogen with an initial rate of 100 TON/h upon irradiation at $\lambda > 450$ nm [68]. Addition of free dmgH$_2$ provides increased stability to the photosystem providing 900 TON relative to the LA (*ca.* 20 mL of hydrogen) in 14 h. A maximum $\Phi = 0.04$ is observed when a bandpass filter at 520 nm is employed, but with a lower TON relative to $\lambda > 450$ nm. The rate of hydrogen evolution shows marked pH dependence with maximum at pH = 7 and decreases sharply at more acidic or basic pH.

Photochemical hydrogen production occurs in a photosystem consisting of [ReBr(CO)$_3$bpy)], [Co(dmgH)$_2$], acetic acid/water, TEOA in DMF at $\lambda > 400$ nm [69]. Turnover numbers are approximately twice compared to when [Ru(bpy)$_3$]$^{2+}$ is the LA under the same conditions (150 TON in 9 h for [ReBr(CO)$_3$bpy)] relative to 65 TON when [Ru(bpy)$_3$]$^{2+}$ is the LA). The quantum yield for hydrogen production is 0.26. The rate of hydrogen production shows a square dependence on the total Co concentration. This square dependence implies a homolytic pathway for hydrogen photocatalysis. The rate of hydrogen production


R=H [Co(dmgH)$_2$pyCl]
R=COOMe [Co(dmgH)$_2$(4-COOMe-py)Cl]
R= NMe$_2$ [Co(dmgH)2(4-Me2N-py)Cl]
[Co(dmgH)(dmgH$_2$)Cl$_2$]
[Co(dmgH)$_2$(py)$_2$]$^+$
[Co(dmgH)$_2$(P(n-Bu)$_3$)Cl]
[Co(dmgBF$_2$)$_2$(OH$_2$)$_2$]
TEOA$^+$
+ glycoladehyde
+ di(ethanol)amine
+ H$^+$ + e$^-$
[Pt-LA]$^-$
TEOA
[*Pt-LA]
hv
[Pt-LA]
H$_2$
Co^{II}
H$^+$
Co^{II}H
Co^{I}
Co^{III}H
e$^-$
H$^+$

Hydrogen via Direct Solar Production. Figure 19
Representative cobaloximes used for solar hydrogen photocatalysis and the proposed photocycle for hydrogen production by reductive quenching of the LA



Hydrogen via Direct Solar Production. Figure 20
General representation of a homolytic pathway using a Co-based hydrogen producing photosystem

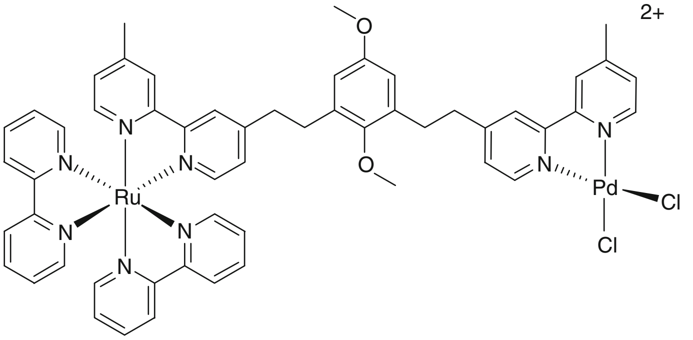
exhibited a linear dependence on the photon flux. Forward electron transfer between the LA and the CAT occurs at a rate close to diffusion control, with a rate of $2.5 \times 10^8\ M^{-1}\ s^{-1}$. An excess of dimethylglyoxime is systematically added to the system to ensure the complete formation of the CAT and reduce the portion of free $[Co]_{solv}{}^{2+}$, which is also an efficient quencher of the excited state of the LA, Fig. 20.

Pt and Pd-Based Homogeneous Systems for Solar Hydrogen Production

Pt- and Pd-based systems have been used in solar hydrogen production schemes. Supramolecular systems that couple polyazine LAs with bridged Pt or Pd CATs for proton reduction are known [70, 73, 74, 101, 102]. The use of the supramolecular design allows for increased electron transfer rates and component modification to enhance device function and light-absorbing properties. The bridged systems use rigid or flexible spacers between the LA and the CAT which controls the level of electronic communication between components. The formation of metallic colloids during photoreduction of the supramolecular assembly has been reported in some studies using similar supramolecular assemblies, implying that heterogeneous catalysis is dominant with the colloidal metal being the active catalytic species [103, 104]. Photoinduced decomposition of $[(bpy)_2Ru(Mebpy(CH_2)_2C_6H_2(OCH_3)_2(CH_2)_2Mebpy)PdCl_2]^{2+}$, Fig. 21, was observed by ESI-MS and the formation of Pd colloid observed by TEM and XPS of evaporated reaction mixtures during photochemical hydrogen production [103]. Mercury tests are employed to deactivate photoproduced metal colloids via formation of inactive alloys with colloidal metals eliminating the impact of the colloid on photochemical activity [86, 87].

These tests can provide meaningful information regarding the mechanism of action for solar hydrogen production, but care must be taken in drawing conclusions as this is a complicated issue. Concentration of the medium to study solids through TEM and XPS can also lead to system decomplexation resulting in formation of metallic solids. Reduced polyazine complexes including those produced electrochemically are known to adsorb to electrodes or metal surfaces such as Hg(l) and this could deactivate a catalytic system. The study of complicated multistep photocatalytic processes is complex, and detailed analyses with carefully selected comparison to determine the catalytically active species are necessary.

A RuPt bimetallic system $[(bpy)_2Ru(phenNHCO(COOHbpy))PtCl_2]^{2+}$, which employs an amide linkage on the phenanthroline as a bridge, produces hydrogen from water when illuminated at $\lambda > 350$ nm in aqueous acetate buffer solutions in the presence of EDTA, with a $\Phi = 0.01$ and 5 turnovers in 10 h [70]. This complex displays an emission from the ^{3}MLCT state at $\lambda_{max}{}^{em} = 610$ nm in an aqueous solution with its intensity quenched by 67% relative to the parent $[(bpy)_2Ru(phenNHCO(COOHbpy))]^{2+}$

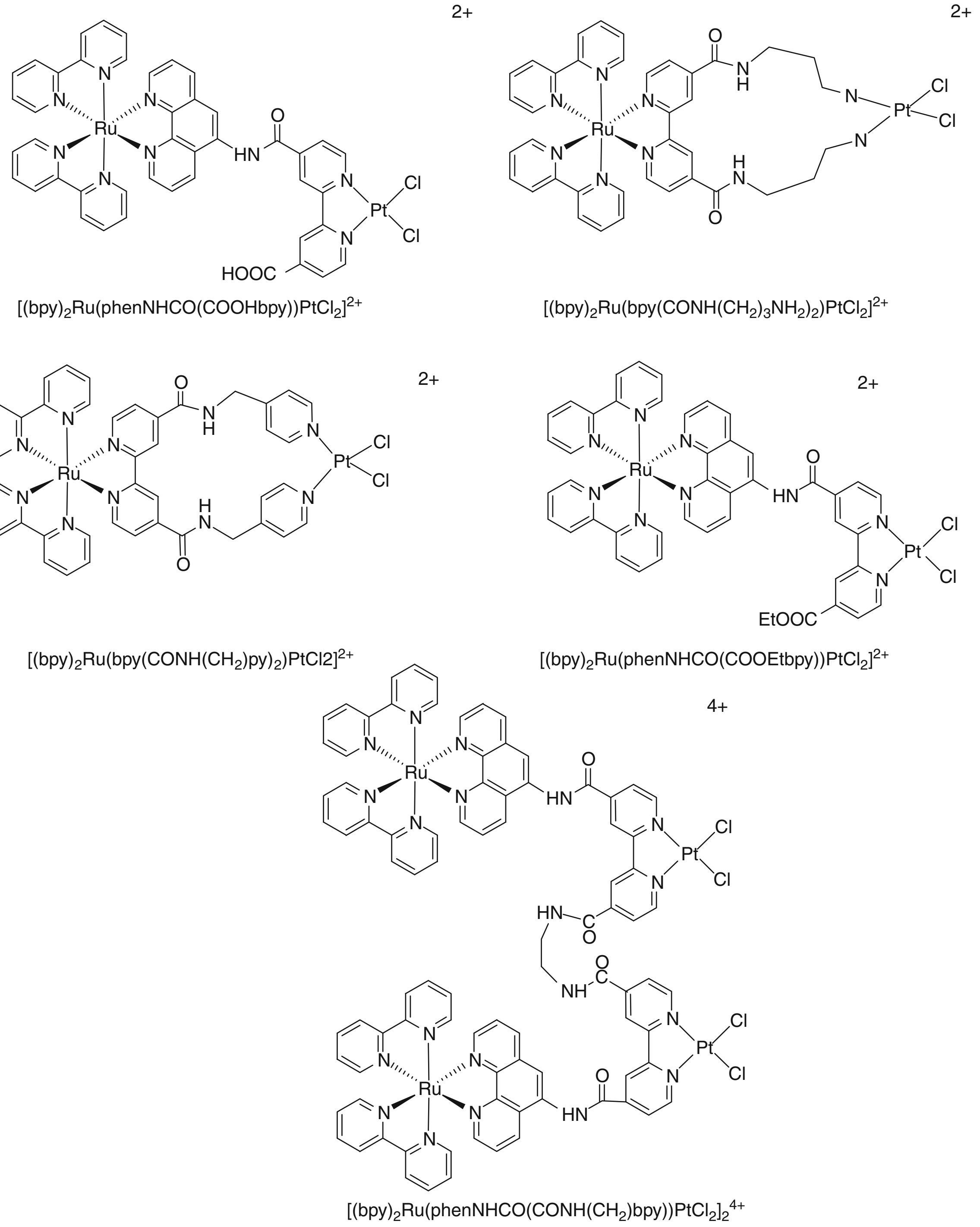


Hydrogen via Direct Solar Production. Figure 21
Representation of $[(bpy)_2Ru(Mebpy(CH_2)_2C_6H_2(OCH_3)_2(CH_2)_2Mebpy)PdCl_2]^{2+}$ (bpy = 2,2′-bipyridine)

suggestive of intramolecular electron transfer process occurring in the bimetallic system. To ascertain that there is no interference from any photodecomposition products, comparisons are made with analogous multicomponent photosystems. The system consisting of $[Ru(bpy)_3](NO_3)_2$, N,N′-dimethyl-4,4′-bipyridinium, and $[PtCl_2(dcbpy)]$ (dcbpy = 4,4′-dicarboxy-2,2′-bipyridine) is effective providing a $\Phi = 0.02$, but completely retarded in the absence of N,N′-dimethyl-4,4′-bipyridinium. Hydrogen production does not occur when $[Ru(bpy)_2(5\text{-aminophen})]^{2+}$ and $[PtCl_2(dcbpy)]$ or the parent LA, $[(bpy)_2Ru(phenNHCO(COOHbpy))]^{2+}$, and K_2PtCl_4 is the photosystem. The lack of activity by these multicomponent systems suggests the bridge between Ru and Pt is essential for photoreactivity in this system. Hydrogen production shows a linear response to photon flux implying that one-photon excitation of a molecule operates. Rate of hydrogen production shows a linear dependence on supramolecular complex concentration. Lower photocatalytic activity is observed when the COOH on bpy is replaced by a COOEt group demonstrating that a clear relationship exists between the energy of the LUMO and the hydrogen evolving activity which was greatest when the more electron withdrawing COOH is the substituent [105].

The hydrogen evolution rate by $[(bpy)_2Ru(phenNHCO(COOHbpy))PtCl_2]^{2+}$ is shown to obey Michaelis-Menten enzymatic kinetics with regard to the concentration of EDTA, suggesting that an ion-pair formation between $[(bpy)_2Ru(phenNHCO(COOHbpy))PtCl_2]^{2+}$ and $EDTA^{2-}$ is an important step in the photocycle [106]. Bimetallic RuPt systems that incorporate aliphatic linkages to bridge $[(bpy)_2Ru(bpy(CONH(CH_2)_3NH_2)_2)PtCl_2]^{2+}$ and $[(bpy)_2Ru(bpy(CONH(CH_2)py)_2)PtCl_2]^{2+}$, Fig. 22, do not display photocatalytic activity. This implies that aromaticity of the bridge connecting the LA and the CAT is important for intramolecular electron transfer in $[(bpy)_2Ru(phenNHCO(COOHbpy))PtCl_2]^{2+}$. Recently a tetrametallic system that shows higher catalytic activity to the original bimetallic system has been reported, $[(bpy)_2Ru(phenNHCO(CONH(CH_2)bpy))PtCl_2]_2{}^{4+}$ [106]. This system shows enhanced catalytic activity with a $\Phi = 0.014$.

The tetrametallic $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ and $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ systems have been investigated as photoinitiated electron collectors (dpq = 2,3-bis(2-pyridyl)quinoxaline), which photocatalyze the production of hydrogen from water with $\Phi = 0.01$ [71, 72]. These systems couple two terminal $(TL)_2Ru^{II}(dpp)$ LA units to a central ruthenium which is connected to a reactive platinum component through a dpq bridge, Fig. 23. $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ displays a Ru $(d\pi) \rightarrow dpp(\pi^*)$ CT absorption observed at 541 nm and an emission from the ^{3}MLCT state ($\lambda_{max}{}^{em}$ = 752 nm and $\Phi^{em} = 3.0 \times 10^{-4}$) with reduced intensity relative to $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)]^{6+}$ ($\lambda_{max}{}^{em}$ = 752 nm and $\Phi^{em} = 12.3 \times 10^{-4}$) in acetonitrile [71]. The τ of $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ is 120 ns, also reduced relative to the $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$, 170 ns. The reductions in the Φ^{em} and τ are characteristic of intramolecular electron transfer in $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$



$[(bpy)_2Ru(phenNHCO(COOHbpy))PtCl_2]^{2+}$

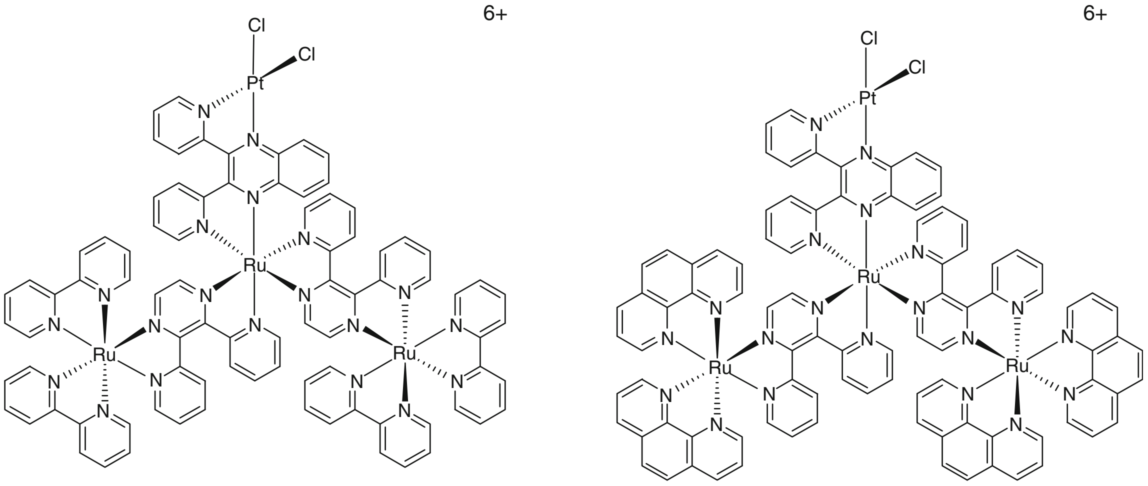
$[(bpy)_2Ru(bpy(CONH(CH_2)_3NH_2)_2)PtCl_2]^{2+}$

$[(bpy)_2Ru(bpy(CONH(CH_2)py)_2)PtCl2]^{2+}$

$[(bpy)_2Ru(phenNHCO(COOEtbpy))PtCl_2]^{2+}$

$[(bpy)_2Ru(phenNHCO(CONH(CH_2)bpy))PtCl_2]_2^{4+}$

Hydrogen via Direct Solar Production. Figure 22
RuPt supramolecular systems for solar energy conversion



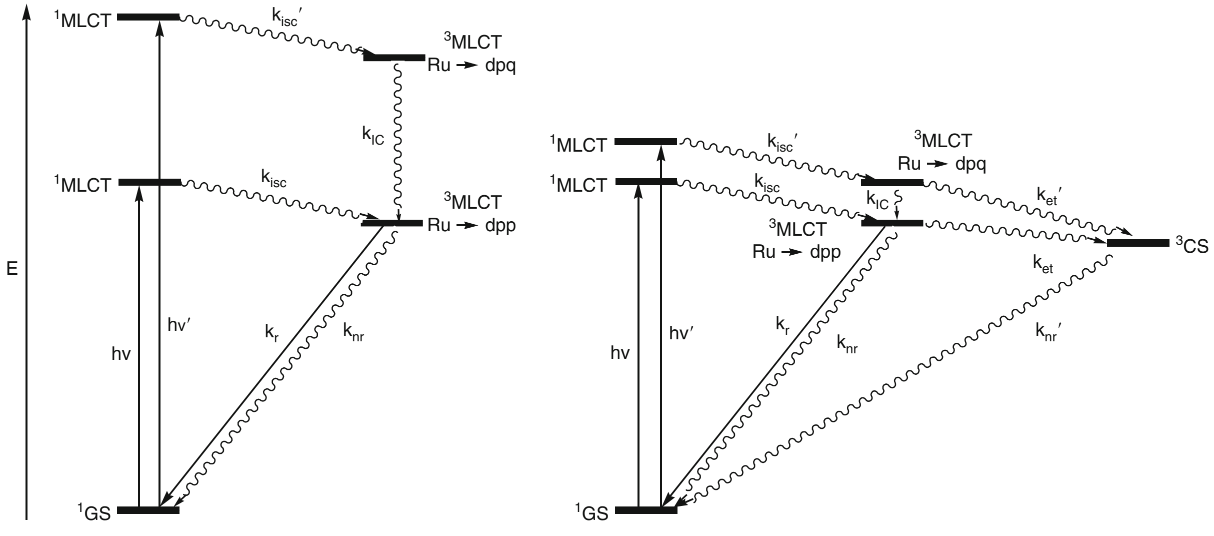
$[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$

Hydrogen via Direct Solar Production. Figure 23
Representations of $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ and $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$

populating a lowest energy CS state with k_{et} = 2.5×10^6 s^{-1}. The τ = 1.7 μs at 77 K in a solid glass matrix in both tetrametallic and trimetallic systems. The emission of $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ is quenched efficiently by DMA with k_q of 4.8×10^9 M^{-1} s^{-1}. The properties of $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ with bpy substituted TL are similar. $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ displays Ru(dπ) → dpp(π*) CT absorption at 540 nm and an emission from the ^{3}MLCT state ($\lambda_{max}{}^{em}$ = 745 nm and $\Phi^{em} = 2.5 \times 10^{-4}$) with reduced intensity relative to $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)]^{6+}$ ($\lambda_{max}{}^{em}$ = 745 nm and $\Phi^{em} = 6.0 \times 10^{-4}$) in acetonitrile [72]. The τ of $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ is 92 ns, also reduced relative to the $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)]^{6+}$, 133 ns. The reductions in the Φ^{em} and τ are characteristic of intramolecular electron transfer in $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ populating a lowest energy CS state with $k_{et} = 3.4 \times 10^6$ s^{-1}. Intramolecular electron transfer is hindered at 77 K in a solid glass matrix with τ = 1.7 μs in both tetrametallic and trimetallic systems. The emission of $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ is quenched efficiently by DMA with k_q of 5.4×10^9 M^{-1} s^{-1}.

A unique feature in these tetrametallic systems is that a greater degree of reduction of the Φ^{em} in the tetrametallic vs. the trimetallic is observed relative to the reduction of the τ of the tetrametallic vs. trimetallic. For $[Ru(bpy)_3]^{2+}$ and similar systems, the $\Phi_{^3MLCT} = 1$. The differences in the degrees of quenching of Φ^{em} and τ of the tetrametallic vs. model trimetallic systems imply that $\Phi_{^3MLCT} \neq 1$, which suggests a reduction in population of the emitting ^{3}MLCT excited state due to an alternative intramolecular electron transfer pathway (k_{et}') in which the higher-lying central Ru(dπ) → dpq(π*) CT state directly populates the ^{3}CS, Fig. 24.

Both $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ and $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ are photoinitiated electron collectors with electrons being collected on the molecular orbitals of the central Ru and the surrounding BLs. Spectroscopic changes that occur when the systems are photochemically reduced in the presence of DMA or electrochemically reduced by two electrons are almost identical, implying that the photoproduct and electrochemically reduced product are the same. For $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$, photochemical hydrogen production occurs in the presence of DMA and water when photoexcited at λ = 530 nm with Φ = 0.01 representing 40 turnovers in 3 h. The amount of hydrogen produced by $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ is much higher than that



$[\{(TL)_2Ru(dpp)\}_2Ru(dpq)](PF_6)_6$ $[\{(TL)_2Ru(dpp)\}_2Ru(dpq)PtCl_2](PF_6)_6$

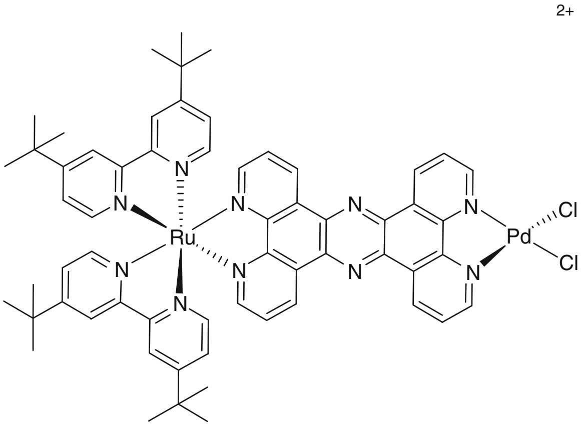
Hydrogen via Direct Solar Production. Figure 24
State diagram showing excited state photophysical properties of $[\{(TL)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ and $[\{(TL)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$ (TL = bpy or phen)

produced by $[\{(bpy)_2Ru(dpp)\}_2Ru(dpq)]^{6+}$ and Pt colloid. For $[\{(phen)_2Ru(dpp)\}_2Ru(dpq)PtCl_2]^{6+}$, photochemical hydrogen production occurs in the presence of DMA when photoexcited at $\lambda = 470$ nm with 115 TON in 5 h. Addition of excess Hg does not impede catalytic activity, which implies that the supramolecular architecture remains intact during the photocycle and colloidal Pt is not operative.

Bifunctional Pt-Based System The complex [PtCl(tpy)]Cl is a single molecule that acts in a bifunctional manner promoting both photosensitization and catalysis [73]. Photolysis of [PtCl(tpy)]Cl in aqueous buffer at $\lambda = 350$–550 nm in the presence of EDTA affords 3 TON in 7 h with a $\Phi = 0.02$. 2-Morpholinoethanesulfonic acid is the buffer used to avoid coordination of buffer reagents to the Pt^{II} by replacing Cl, shown to lower hydrogen yield. Contribution from catalytically active Pt colloids seems not operative as the system functions with added Hg(l). It is suggested that photochemical hydrogen production proceeds through a bimolecular mechanism. Photosensitization is suggested to occur through a ^{3}MMLCT state (MMLCT = metal-metal-to-ligand charge-transfer) in dimer formation facilitated by Pt····Pt association. The saturation kinetics with regard to the EDTA concentration suggests that the dimer of $[PtCl(tpy)]^+$ and the dianionic form of EDTA form an ion-pair adduct to facilitate reductive quenching by EDTA during the photochemical processes.

Palladium-Based Supramolecular Solar Hydrogen Photocatalysts

The RuPd bimetallic system $[(^tBu_2bpy)_2Ru(tpphz)PdCl_2]^{2+}$ (^{t}Bu = tertiary butyl, tpphz = tetrapyridophenazine, Fig. 25) is a solar hydrogen photocatalyst producing hydrogen, with 56 TON in 30 h, in the presence of the electron donor, Et_3N, in acetonitrile when optically excited at $\lambda = 470$ nm [74]. The proton source is Et_3N. A weak emission at $\lambda_{max}{}^{em} = 650$ nm with $\tau = 27$ ns is observed significantly quenched relative to the parent $[(^tBu_2bpy)_2Ru(tpphz)]^{2+}$ ($\lambda_{max}{}^{em} = 638$ nm, $\tau = 154$ ns). It is suggested that the Ru^{II}-phenazine$^-$-Pd^{II} radical photogenerated in the presence of Et_3N is involved in the photochemical hydrogen generation step, initiating chloride loss from palladium. Photoinduced electron transfer



Hydrogen via Direct Solar Production. Figure 25 Representation of $[(^{t}Bu_2bpy)_2Ru(tpphz)PdCl_2]^{2+}$

processes to reduce the Pd site have been studied by resonance Raman and ultrafast time-resolved spectroscopy showing rapid Pd reduction [107].

Conclusions and Future Directions

The conversion of solar energy to a transportable fuel is challenging but important in meeting increasing energy demands. The production of hydrogen fuel via solar water splitting is an attractive approach to alternative energy conversion and storage as the amount of energy reaching the Earth's surface from the sun in 1 h is sufficient to power our planet for more than a year. Photocatalysis to reduce water to produce hydrogen is complicated involving multielectron transfer, bond breaking, and bond formation. Despite these challenges and the lack of consistent substantial long-term funding, a number of photosystems for solar hydrogen production have been developed recently. Pioneering work on multicomponent systems was critical to success. Water splitting to produce hydrogen is complicated and involves energetically demanding chemical reactions with bond breaking and formation and is a multielectron process. Solar energy conversion is promoted via water reduction by multistep processes with multiple photons used sequentially.

Work has focused on the LA and/or CAT design for solar hydrogen production schemes. Structural diversity is starting to be available. Many systems incorporate Ru-based LAs and reactive metal centers in the photocatalytic schemes. Ir, Pt, Re, and Zn-based systems have emerged recently showing promise as LAs. Biomimetic iron-based systems, Rh, Co, Pt, and Pd-based CATs have been typically used is these schemes. The Rh- and Co-based systems take advantage of ability of Rh or Co to cycle through three oxidation states reducing from M^{III} to M^{II} and M^{I} for multielectron chemistry. The application of typical Ru (II) trischelate systems with ligands such as bpy or phen provide for absorption in the visible region to *ca.* 450 nm. The light absorption energy can be modulated by ligand substitution and the use of ligands with stabilized π^* orbitals provide for longer wavelength excitation often to *ca.* 550–650 nm.

Using bridging ligands provides a means to covalently attach the LA to the CAT allowing more efficient intramolecular electron transfer pathways from the LA to the catalytic site. In addition, factors that impact intramolecular electron transfer can be studied and modulated within molecular systems. Coupling of a LA to CAT often dramatically perturbs the properties of these components. The ability to modulate the electrochemical and photochemical properties by component modification in supramolecular systems can be exploited in molecular design. The application of large molecular systems to address hydrogen photocatalysis allows system study using conventional molecular characterization methods allowing for the ability to probe the processes involved in the photocatalytic process.

Many factors are considered during photosystem design for hydrogen photocatalysis. The redox potentials of the LA, CAT, and coordinated ligands dictate the driving force of electron transfer reactions. LA properties including absorptivity and excited state lifetimes are contributing factors which determine wavelength response and overall photosystem efficiency. System parameters including, solvent, electron donor, pH, wavelength of excitation, etc., also determine functioning system efficiency and stability. Often solvent choice can have dramatic impact with systems showing good photoreactivity in organic solvents that are not functional in aqueous media. Catalysts and electron donors selected should not compete for light with the LA. Study and comparison of system parameters that impact hydrogen photocatalysis is crucial as many conditions impact photocatalysis and no set of standard conditions exist. For example, some solvents that show enhanced system function for photocatalytic systems have been shown to deactivate other systems. Comparison of systems is complicated by these widely varied conditions.

The study of electron transfer between components is key to understanding photosystem properties. The electron transfer events occurring in supramolecular systems are studied with comparison using model systems that should have similar spectroscopic and photophysical properties, but is always not the case. Careful selection of model systems is essential to these studies as the assumption that k_r and k_{nr} are the same in the model systems and supramolecular systems is inherent to this analysis. Sample purity is also important in these studies as the rate of electron transfer could be underestimated in the presence of other emissive impurities or overestimated in the presence of impurities which quench the emissive excited state. The study of these electron transfer events become more complicated as model systems to probe excited state processes are not readily available.

Design and development of functioning photocatalysts for the reduction of water to produce hydrogen fuel has become an international effort as the world recognizes the significance for the need for environmentally friendly, renewable, economically viable alternatives for fuel production. There has been much success in this field given the complexity of the systems and processes.

The iron-based systems exploit the ability of iron-based hydrogenases to catalyze water reduction to produce hydrogen. Interesting biomimetic assemblies have been produced that couple ruthenium to diiron catalytic centers but these complexes have not shown applicability to solar hydrogen photocatalysis. The unfavorable thermodynamic driving force for electron transfer from the ruthenium to the CAT impedes the use of the supramolecules diiron systems as photocatalysts for solar hydrogen production. The coupling of alternative ZnTPP LA units to these diiron cores provides for thermodynamically favorable electron transfer and has provided for photocatalytic systems with low TON. The ZnTPP LA unit possesses a higher energy singlet excited state that allows electron transfer to the diiron unit. These supramolecular assemblies reduce acidic solutions of 2-mercaptobenzoic acid to produce hydrogen following optical excitation with low turnovers. The multicomponent systems using diiron catalytic centers and Ru-based LA have shown promise providing higher TON using high CAT loading. In these systems, the stability of the CAT determines system stability with system malfunction occurring often after few hours due to CAT decomposition. Use of Ir-based LAs have shown better promise with higher TON relative to the LA and CAT, but system deactivation occurs due to decomplexation of the CAT.

Rhodium-centered photoinitiated electron collectors are known to catalyze water reduction to hydrogen. $[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$, $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$, and $[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$ are the lead photocatalysts in this area, demonstrating higher hydrogen yields and turnover capacities. The use of a bridging dpp on the Ru LA provides significant coverage of the solar spectrum to the low energy visible. The Ru LA displays an emissive ^{3}MLCT excited state. The spectroscopic and photophysical properties display population of the emissive ^{3}MLCT excited state followed by intramolecular electron transfer to populate the ^{3}MMCT excited state. Optical excitation in the presence of an electron donor leads to at least two intramolecular electron transfer steps affording a coordinatively unsaturated Rh^I species which can interact with and reduce a substrate such as water. These systems couple two charge-transfer light absorbers to a single electron collector. Studies on the design considerations for a functioning system for optical multielectron reduction of water to produce hydrogen have shown that this general structural motif is applicable in solar hydrogen production schemes. A related photoinitiated electron collector, $[\{(bpy)_2Ru(dpb)\}_2IrCl_2]^{5+}$, does not demonstrate photocatalytic activity implying that the Rh center is necessary and facilitates the chemical steps involved in bond transformations. The coordination environment on the rhodium center impacts the photocatalytic activity as evidenced by the greater hydrogen yields when weaker σ-donors are present on rhodium (Br vs. Cl), further implying that the reactive rhodium center and/or the ^{3}MMCT excited state may be important. $[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$ provides a system with a lower lying $Rh(d\sigma^*)$ acceptor orbital with a larger driving force for intramolecular electron transfer to produce the ^{3}MMCT state and/or promotes halide loss to generate the Rh^{I} system. Systematic studies on factors impacting hydrogen production, including the driving force for reductive quenching of the excited states by the electron donor, excited state lifetime of the photocatalyst, effective pH of the solution, and electron donor-catalyst interactions, has been studied. Photocatalytic activity of $[\{(tpy)RuCl(dpp)\}_2RhCl_2]^{3+}$ and $[\{(bpy)_2Os(dpp)\}_2RhCl_2]^{5+}$ is observed with lower efficiency, consistent with the lower driving force for reductive quenching of the ^{3}MLCT excited state by DMA. The result is important, showing hydrogen can be evolved employing the ^{3}MLCT excited state as electron transfer quenching of the ^{3}MMCT state is not thermodynamically favorable in these systems. The system pH and the electron donor-catalyst interactions are important factors to be considered as varied photocatalytic activity is observed with variation of electron donors.

Multicomponent systems incorporating cyclometalated Ir complexes have been used as LAs in photochemical schemes for production of hydrogen from water using $[Rh(NN)_3]^{3+}$ CATs. A preliminary study of a library of LA and CAT has enabled the identification of the most efficient system providing high TON and good system stability. Ir LAs provide varied excited state lifetimes and redox potentials relative to the classical Ru LAs.

The use of cobalt CATs have been investigated in solar hydrogen production schemes. The often substitutionally labile cobalt centers are stabilized by the use of cobaloximes often with excess ligand in solution. The supramolecular systems incorporating Ir-based LAs show superior activity to the Ru-based system and enhanced stability relative to the analogous multicomponent system. The use of a Re-based LA with cobaloxime CAT has provided good functioning. Use of Pt- or Ir-based LAs with cobaloxime CATs has shown to be the most promising, providing high TON and good system stability.

Recent studies have focused on the coupling of Pd and Pt reactive metal centers to Ru charge-transfer light absorbers for multielectron solar hydrogen photocatalysis. There have been conflicting reports on these systems with follow-up studies showing that in some RuPt and RuPd systems the reduction of the complex may lead to colloidal Pd or Pt. It was proposed that Pt(s) or Pd(s) are the active catalyst for hydrogen generation. Some systems are proposed to function without colloidal metal formation. Careful analysis of these catalytic systems is needed to provide insight into the function of these photocatalysts and the optimization of the future generation of catalysts.

The area of solar hydrogen production through the use of photocatalysts that collect and deliver reducing equivalents is an active area of current research in inorganic photochemistry. The design and development of systems that catalyze this complicated process can play a significant role in the development of efficient photocatalysts for solar hydrogen production. The fundamental understanding of the properties of the LA, bridging ligand, CAT, and system parameters that determine hydrogen production efficiency is critical to the successful development of efficient and stable photocatalysts for the reduction of water to produce hydrogen fuel. The development of structurally diverse systems which couple a LA to a CAT is needed in this field with systematic variation of device components. Basic chemical studies of the perturbation of properties upon assembly of the supramolecules are also needed. Careful analysis of results and comparisons are needed. It is important that long-term studies of these important systems for light to fuel conversion are undertaken to understand the functioning of these systems and explore the multidimensional parameter space needed to optimize the performance of these systems that are still in their infancy. Detailed mechanistic studies are necessary to uncover factors that limit the functioning of these systems and design of more promising molecular architectures that carry out this complicated photocatalytic reaction. A wide array of system parameters can impact system functioning further complicating this process. The research and development in this very important area requires long-term interest and research funding to assemble the fundamental chemical knowledge needed to establish efficient solar hydrogen production schemes. Sustained and focused research is needed to bring these systems to completion.

Acknowledgments

Acknowledgment is made to all the students and research scientists who have worked in this area in the Brewer Group. Special acknowledgment to Ms. Jessica D. Knoll for help with the manuscript. Acknowledgment is also made to the Chemical Sciences, Geosciences and Biosciences Division, Office of Basic Energy Sciences, Office of Sciences, US Department of Energy for their generous support of our research and to the financial collaboration of Phoenix Canada Oil Company which holds long-term license rights to commercialize our Rh-based technology.

Bibliography

1. Bard AJ, Fox MA (1995) Artificial photosynthesis: solar splitting of water to hydrogen and oxygen. Acc Chem Res 28:141–145
2. DOE (2005) Report of the basic energy sciences workshop on solar utilization. Department of Energy, Washington, DC
3. Lewis NS (2009) A perspective on forward research and development paths for cost-effective solar energy utilization. ChemSusChem 2:383–386
4. Nocera DG (2009) Living healthy on a dying planet. Chem Soc Rev 38:13–15
5. Bockris JO'M (1999) Hydrogen economy in the future. Int J Hydrogen Energ 24:1–15
6. Lubitz W, Tumas W (2007) Hydrogen: an overview. Chem Rev 107:3900–3903
7. http://www.fsec.ucf.edu/EN/consumer/hydrogen/basics/production.htm
8. Balzani V, Moggi L, Scandola F (1987) Towards a supramolecular photochemistry: assembly of molecular components to obtain photochemical molecular devices. In: Balzani V (ed) Supramolecular photochemistry. Reidel, Dordrecht, pp 1–28
9. Kalyanasundaram K (1982) Photophysics, photochemistry and solar energy conversion with tris(bipyridyl)ruthenium(II) and its analogues. Coord Chem Rev 46:159–244
10. Juris A, Balzani V, Barigelletti F, Campagna S, Belser P, Von Zelewsky A (1988) Ru(II) polypyridine complexes: photophysics, photochemistry, eletrochemistry, and chemiluminescence. Coord Chem Rev 84:85–277
11. Kirch M, Lehn J-M, Sauvage J-P (1979) Hydrogen generation by visible light irradiation of aqueous solutions of metal complexes. An approach to the photochemical conversion and storage of solar energy. Helv Chim Acta 62:1345–1384
12. Balzani V, Juris A, Venturi M, Campagna S, Serroni S (1996) Luminescent and redox-active polynuclear transition metal complexes. Chem Rev 96:759–834
13. Kawanishi Y, Kitamura N, Tazuke S (1989) Dependence of spectroscopic, electrochemical, and excited-state properties of tris chelate ruthenium(II) complexes on ligand structure. Inorg Chem 28:2968–2975
14. Durham B, Caspar JV, Nagle JK, Meyer TJ (1982) Photochemistry of $Ru(bpy)_3^{2+}$. J Am Chem Soc 104:4803–4810
15. Demas JN, Adamson AW (1971) A new photosensitizer. Tris (2,2′-bipyridine)ruthenium(II) chloride. J Am Chem Soc 93:1800–1801
16. Gafney HD, Adamson AW (1972) Excited state $Ru(bipyr)_3^{2+}$ as an electron-transfer reductant. J Am Chem Soc 94:8238–8239
17. Bock CR, Meyer TJ, Whitten DG (1974) Electron transfer quenching of the luminescent excited state of tris(2,2′-bipyridine)ruthenium(II). Flash photolysis relaxation technique for measuring the rates of very rapid electron transfer reactions. J Am Chem Soc 96:4710–4712
18. Bock CR, Connor JA, Gutierrez AR, Meyer TJ, Whitten DG, Sullivan BP, Nagle JK (1979) Estimation of excited-state redox potentials by electron-transfer quenching. Application of electron-transfer theory to excited-state redox processes. J Am Chem Soc 101:4815–4824
19. Tinker LL, McDaniel ND, Bernhard S (2009) Progress towards solar-powered homogeneous water photolysis. J Mater Chem 19:3328–3337
20. Wang M, Na Y, Gorlov M, Sun L (2009) Light-driven hydrogen production catalysed by transition metal complexes in homogeneous systems. Dalton Trans 33:6458–6467
21. Inagaki A, Akita M (2010) Visible-light promoted bimetallic catalysis. Coord Chem Rev 254:1220–1239
22. Balzani V, Credi A, Venturi M (2008) Photochemical conversion of solar energy. ChemSusChem 1:26–58
23. Lin CT, Böttcher W, Chou M, Creutz C, Sutin N (1976) Mechanism of the quenching of the emission of substituted polypyridineruthenium(II) complexes by iron (III), chromium(III), and europium(III) ions. J Am Chem Soc 98:6536–6544
24. Kober EM, Sullivan BP, Dressick WJ, Caspar JV, Meyer TJ (1980) Highly luminescent polypyridyl complexes of osmium (II). J Am Chem Soc 102:7383–7385
25. Braunstein CH, Baker AD, Strekas TC, Gafney HD (1984) Spectroscopic and electrochemical properties of the dimer tetrakis (2,2′-bipyridine)(μ-2,3-bis(2-pyridyl)pyrazine)diruthenium(II) and its monomeric analogue. Inorg Chem 23:857–864
26. Fuchs Y, Lofters S, Dieter T, Shi W, Morgan R, Strekas TC, Gafney HD, Baker AD (1987) Spectroscopic and electrochemical properties of dimeric ruthenium(II) diimine complexes and determination of their excited state redox properties. J Am Chem Soc 109:2691–2697
27. Wallace AW, Murphy WR, Petersen JD (1989) Electrochemical and photophysical properties of mono- and bimetallic ruthenium(II) complexes. Inorg Chim Acta 166:47–54
28. Richter MM, Brewer KJ (1991) Synthesis and characterization of osmium(II) complexes incorporating polypyridyl bridging ligands. Inorg Chim Acta 180:125–131
29. Abdel-Shafi AA, Worrall DR, Ershov AY (2004) Photosensitized generation of singlet oxygen from ruthenium(II) and osmium (II) bipyridyl complexes. Dalton Trans 7:30–36
30. Winkler JR, Netzel TL, Creutz C, Sutin N (1987) Direct observation of metal-to-ligand charge-transfer (MLCT) excited states of pentaammineruthenium(II) complexes. J Am Chem Soc 109:2381–2392
31. Beley M, Collin J-P, Sauvage J-P, Sugihara H, Heisel F, Miehé A (1991) Photophysical and photochemical properties of ruthenium and osmium complexes with substituted terpyridines. J Chem Soc Dalton Trans 23: 3157–3159

32. Rillema DP, Mack KB (1982) The low-lying excited state in ligand π-acceptor complexes of ruthenium(II): mononuclear and binuclear species. Inorg Chem 21:3849–3854
33. Berger RM (1990) Excited-state absorption spectroscopy and spectroelectrochemistry of tetrakis(2,2′-bipyridine)(μ-2,3-bis (2-pyridyl)pyrazine)diruthenium(II) and its Monomeric counterpart: a comparative study. Inorg Chem 29:1920–1924
34. Balzani V, Campagna S, Denti G, Juris A, Serroni S, Venturi M (1998) Designing dendrimers based on transition-metal complexes: light-harvesting properties and predetermined redox patterns. Acc Chem Res 31:26–34
35. Campagna S, Pietro CD, Loiseau F, Maubert B, McClenaghan N, Passalacqua R, Puntoriero F, Ricevuto V, Serroni S (2002) Recent advances in luminescent polymetallic dendrimers containing the 2,3-Bis(2′-pyridyl)pyrazine bridging ligand. Coord Chem Rev 229:67–74
36. Molnar SM, Nallas GNA, Bridgewater JS, Brewer KJ (1994) Photoinitiated electron collection in a mixed-metal trimetallic complex of the form $[\{(bpy)_2Ru(dpb)\}_2IrCl_2](PF_6)_5$ (bpy = 2,2′-bipyridine and dpb = 2,3-bis(2-pyridyl)benzoquinoxaline). J Am Chem Soc 116:5206–5210
37. Konduri R, Ye H, MacDonnell FM, Serroni S, Campagna S, Rajeshwar K (2002) Ruthenium photocatalysts capable of reversibly storing up to four electrons in a single acceptor ligand: a step closer to artificial photosynthesis. Angew Chem Int Edit 41:3185–3187
38. Konduri R, de Tacconi NR, Rajeshwar K, MacDonnell FM (2004) Multielectron photoreduction of a bridged ruthenium dimer, $[(phen)_2Ru(tatpp)Ru(phen)_2](PF_6)_4$: aqueous reactivity and chemical and spectroelectrochemical identification of the photoproducts. J Am Chem Soc 126:11621
39. Polyansky DE, Cabelli D, Muckerman JT, Fukushima T, Tanaka K, Fujita E (2008) Mechanism of hydride donor generation using a Ru(II) complex containing an NAD^+ model ligand: pulse and steady-state radiolysis studies. Inorg Chem 47:3958–3968
40. Fukushima T, Fujita E, Muckerman JT, Polyansky DE, Wada T, Tanaka K (2009) Photochemical stereospecific hydrogenation of a Ru complex with an $NAD^+/NADH^-$Type ligand. Inorg Chem 48:11510–11512
41. Lomoth R, Ott S (2009) Introducing a dark reaction to photochemistry: photocatalytic hydrogen from [FeFe] hydrogenase active site model complexes. Dalton Trans 7:9952–9959
42. Frey M (2002) Hydrogenases: hydrogen-activating enzymes. Chembiochem 3:153–160
43. Capon J-F, Gloaguen F, Pétillon FY, Schollhammer P, Talarmin J (2009) Electron and proton transfers at diiron dithiolate sites relevant to the catalysis of proton reduction by the [FeFe]-hydrogenases. Coord Chem Rev 253:1476–1494
44. Wolpher H, Borgström M, Hammarström L, Bergquist J, Sundström V, Styring S, Sun L, Åkermark B (2003) Synthesis and properties of an iron hydrogenase active site model linked to a ruthenium tris-bipyridine photosensitizer. Inorg Chem Commun 6:989–991
45. Ott S, Kritikos M, Åkermark B, Sun L (2003) Synthesis and structure of a biomimetic model of the iron hydrogenase active site covalently linked to a ruthenium photosensitizer. Angew Chem Int Edit 42:3285–3288
46. Ott S, Borgström M, Kritikos M, Lomoth R, Bergquist J, Åkermark B, Hammarström L, Sun L (2004) Model of the iron hydrogenase active site covalently linked to a ruthenium photosensitizer: synthesis and photophysical properties. Inorg Chem 43:4683–4692
47. Ekström J, Abrahamsson M, Olson C, Bergquist J, Kaynak FB, Eriksson L, Sun L, Becker H-C, Åkermark B, Hammarström L, Ott S (2006) Bio-inspired, side-on attachment of a ruthenium photosensitizer to an iron hydrogenase active site model. Dalton Trans 4599–4606
48. Li X, Wang M, Zhang S, Pan J, Na Y, Liu J, Åkermark B, Sun L (2008) Noncovalent assembly of a metalloporphyrin and an iron hydrogenase active-site model: photo-induced electron transfer and hydrogen generation. J Phys Chem B 112:8198–8202
49. Chong D, Georgakaki IP, Mejia-Rodriguez R, Sanabria-Chinchilla J, Soriaga MP, Darensbourg MY (2003) Electrocatalysis of hydrogen production by active site analogues of the iron hydrogenase enzyme: structure/function relationships. Dalton Trans 4158–4163
50. Lubner CE, Grimme R, Bryant DA, Goldbeck JH (2010) Wiring photosystem I for direct solar hydrogen production. Biochemistry 49:404–414
51. Na Y, Wang M, Pan J, Zhang P, Åkermark B, Sun L (2008) Visible light-driven electron transfer and hydrogen generation catalyzed by bioinspired [2Fe2S] complexes. Inorg Chem 47:2805–2810
52. Streich D, Astuti Y, Orlandi M, Schwartz L, Lomoth R, Hammarström L, Ott S (2010) High-turnover photochemical hydrogen production catalyzed by a model complex of the [FeFe]-hydrogenase active site. Chem Eur J 16:60–63
53. Zhang P, Wang M, Na Y, Li X, Jiang Y, Sun L (2010) Homogeneous photocatalytic production of hydrogen from water by a bioinspired $[Fe_2S_2]$ catalyst with high turnover numbers. Dalton Trans 39:1204–1206
54. Elvington M, Brown J, Arachchige SM, Brewer KJ (2007) Photocatalytic hydrogen production from water employing a Ru, Rh, Ru molecular device for photoinitiated electron collection. J Am Chem Soc 129:10644–10645
55. Arachchige SM, Brown J, Brewer KJ (2008) Photochemical hydrogen production from water using the new photocatalyst $[\{(bpy)_2Ru(dpp)\}_2RhBr_2](PF_6)_5$. J Photochem Photobiol A 197:13–17
56. Arachchige SM, Brown JR, Chang E, Jain A, Zigler DF, Rangan K, Brewer KJ (2009) Design considerations for a system for photocatalytic hydrogen production from water employing mixed-metal photochemical molecular devices for photoinitiated electron collection. Inorg Chem 48:1989–2000
57. Rangan K, Arachchige SM, Brown JR, Brewer KJ (2009) Solar energy conversion using photochemical molecular devices: photocatalytic hydrogen production from water using mixed-metal supramolecular complexes. J Energ Environ Sci 2:410–419

58. Miyake Y, Nakajima K, Sasaki K, Saito R, Nakanishi H, Nishibayashi Y (2009) Design and synthesis of diphosphine ligands bearing an osmium(II) bis(terpyridyl) moiety as a light-harvesting unit: application to photocatalytic production of dihydrogen. Organometallics 28:5240–5243
59. Heyday AF, Nocera DG (2001) Hydrogen produced from hydrohalic acid solutions by a two-electron mixed-valence photocatalyst. Science 293:1639–1641
60. Brown GM, Chan SF, Creutz C, Schwarz HA, Sutin N (1979) Mechanism of the formation of dihydrogen from the photo-induced reactions of tris(bipyridine)ruthenium(II) with tris (bipyridine)rhodium(III). J Am Chem Soc 101:7638–7640
61. Cline ED, Adamson SE, Bernhard S (2008) Homogeneous catalytic system for photoinduced hydrogen production utilizing iridium and rhodium complexes. Inorg Chem 47:10378–10388
62. Fihri A, Artero V, Razavet M, Baffert C, Leibl W, Fontecave M (2008) Cobaloxime-based photocatalytic devices for hydrogen production. Angew Chem Int Edit 47:564–567
63. Fihri A, Artero V, Pereira A, Fontecave M (2008) Efficient H_2-producing photocatalytic systems based on cyclometalated iridium- and tricarbonylrhenium-diimine photosensitizers and cobaloxime catalysts. Dalton Trans 5567–5569
64. Li C, Wang M, Pan J, Zhang P, Zhang R, Sun L (2009) Photochemical hydrogen production catalyzed by polypyridyl ruthenium-cobaloxime heterobinuclear complexes with different bridges. J Organomet Chem 694:2814–2819
65. Brown GM, Brunschwig BS, Creutz C, Endicott JF, Sutin N (1979) Homogeneous catalysis of the photoreduction of water by visible light. Mediation by a tris(2,2′-bipyridine)ruthenium(II)-cobalt(II) macrocycle system. J Am Chem Soc 101:1298–1300
66. Goldsmith JI, Hudson WR, Lowry MS, Anderson TH, Bernhard S (2005) Discovery and high-throughput screening of heteroleptic iridium complexes for photoinduced hydrogen production. J Am Chem Soc 127:7502–7510
67. Du P, Schneider J, Luo G, Brennessel WW, Eisenberg R (2009) Visible light-driven hydrogen production from aqueous protons catalyzed by molecular cobaloxime catalysts. Inorg Chem 48:4952–4962
68. Lazarides T, McCormick T, Du P, Luo G, Lindley B, Eisenberg R (2009) Making hydrogen from water using a homogeneous system without noble metals. J Am Chem Soc 131:9192–9194
69. Probst B, Kolano C, Hamm P, Alberto R (2009) An efficient homogeneous intermolecular rhenium-based photocatalytic system for the production of H_2. Inorg Chem 48:1836–1843
70. Ozawa H, Haga M-A, Sakai K (2006) A photo-hydrogen-evolving molecular device driving visible-light-induced EDTA-reduction of water into molecular hydrogen. J Am Chem Soc 128:4926–4927
71. Knoll JD, Arachchige SM, Brewer KJ (2011) A structurally diverse Ru(II), Pt(II) tetrametallic motif for photoinitiated electron collection and photocatalytic hydrogen production. ChemSusChem 4:252–261
72. Miao R, Zigler DF, Brewer KJ Manuscript to be submitted to Inorg Chem
73. Okazaki R, Masaoka S, Sakai K (2009) Photo-hydrogen-evolving activity of chloro(terpyridine)platinum(II): a single-component molecular photocatalyst. Dalton Trans 6127–6133
74. Rau S, Schäfer B, Gleich D, Anders E, Rudolph M, Friedrich M, Görls H, Henry W, Vos JG (2006) A supramolecular photocatalyst for the production of hydrogen and the selective hydrogenation of tolane. Angew Chem Int Edit 45:6215–6218
75. Schwartz L, Singh PS, Eriksson L, Lomoth R, Ott S (2008) Tuning the electronic properties of $Fe_2(\mu\text{-arenedithiolate})(CO)_{6-n}(PMe_3)_n$ (n = 0, 2) complexes related to the [Fe–Fe]-hydrogenase active site. CR Chim 11:875–889
76. Sutin N, Creutz C, Fujita E (1997) Photo-induced generation of dihydrogen and reduction of carbon dioxide using transition metal complexes. Comment Inorg Chem 19:67–92
77. Balzani V, Credi A, Venturi M (1998) Photochemistry and photophysics of coordination compounds: an extended view. Coord Chem Rev 171:3–16
78. Venturi M, Credi A, Balzani V (1999) Electrochemistry of coordination compounds: an extended view. Coord Chem Rev 185:233–256
79. Scandola F, Argazzi R, Bignozzi CA, Indelli MT (1994) Photoinduced energy and electron transfer in inorganic covalently linked systems. J Photochem Photobiol A 82:191–202
80. Serroni S, Juris A, Campagna S, Venturi M, Denti G, Balzani V (1994) Tetranuclear bimetallic complexes of ruthenium, osmium, rhodium and iridium. Synthesis, absorption spectra, luminescence and electrochemical properties. J Am Chem Soc 116:9086–9091
81. Indelli MT, Scandola F, Collin J-P, Sauvage J-P, Sour A (1996) Photoinduced electron and energy transfer in rigidly bridged Ru(II)-Rh(III) binuclear complexes. Inorg Chem 35:303–312
82. Elvington M, Brewer K (2006) Photoinitiated electron collection at a metal in a rhodium-centered mixed-metal supramolecular complex. Inorg Chem 45:5242–5244
83. White TA, Rangan K, Brewer KJ (2010) Synthesis, characterization, and study of the photophysics and photocatalytic properties of the photoinitiated electron collector $[\{(phen)_2Ru(dpp)\}_2RhBr_2](PF_6)_5$. J Photochem Photobiol A 209: 203–209
84. Kalyanasundaram K, Nazeeruddin MK (1990) Photophysics and photoredox reactions of ligand-bridged binuclear polypyridyl complexes of ruthenium(II) and of their monomeric analogs. Inorg Chem 29:1888–1897
85. Kew G, DeArmond K, Hanck K (1974) Electrochemistry of rhodium-dipyridyl complexes. J Phys Chem 78:727–734
86. Anton DR, Crabtree RH (1983) Dibenzo[a, e]cyclooctatetraene in a proposed test for heterogeneity in catalysts formed from soluble platinum-group metal complexes. Organometallics 2:855–859
87. Baba R, Nakabayashi S, Fujishima A, Honda K (1985) Investigation of the mechanism of hydrogen evolution during photocatalytic water decomposition on metal-loaded semiconductor powders. J Phys Chem 89:1902–1905

88. Arachchige SM, Shaw R, White TA, Shenoy V, Tsui H-M, Brewer KJ (2011) Modification of system parameters for high turnover in a photocatalytic system for water reduction to produce hydrogen using a Ru, Rh, Ru photoinitiated electron collector. ChemSusChem 4:514–518
89. Esswein AJ, Veige AS, Nocera DG (2005) A photocycle for hydrogen production from two-electron mixed-valence complexes. J Am Chem Soc 127:16641–16651
90. Esswein AJ, Dempsey JL, Nocera DG (2007) A $Rh^{II} - Au^{II}$ bimetallic core with a direct metal-metal bond. Inorg Chem 46:2362–2364
91. Cook TR, Esswein AJ, Nocera DG (2007) Metal − halide bond photoactivation from a $Pt^{III}Au^{II}$ complex. J Am Chem Soc 129:10094–10095
92. Chou M, Creutz C, Mahajan D, Sutin N, Zipp AP (1982) Nature of bis(2,2′-bipyridine)rhodium(I) in aqueous solutions. Inorg Chem 21:3989–3997
93. Schwarz HA, Creutz C (1983) Reactions of tris- and bis(2,2′-bipyridine)rhodium(II) complexes in aqueous solution. Inorg Chem 22:707–713
94. Fujita E, Brunschwig BS, Creutz C, Muckerman JT, Sutin N, Szalda D, van Eldik R (2006) Transition state characterization for the reversible binding of dihydrogen to Bis(2,2′-bipyridine) rhodium(I) from temperature- and pressure-dependent experimental and theoretical studies. Inorg Chem 45:1595–1603
95. Dempsey JL, Brunschwig BS, Winkler JR, Gray HB (2009) Hydrogen evolution catalyzed by cobaloximes. Acc Chem Res 42:1995–2004
96. Kölle U (1992) Transition metal-catalyzed proton reduction. New J Chem 16:157–169
97. Hawecker J, Lehn JM, Ziessel R (1983) Efficient homogeneous photochemical hydrogen generation and water reduction mediated by cobaloxime or macrocyclic cobalt complexes. Nouv J Chim 7:271–7
98. Lowry MS, Goldsmith JI, Slinker JD, Rohl R, Pascal RA, Malliaras GG, Bernhard S (2005) Single-layer electroluminescent devices and photoinduced hydrogen production from an ionic iridium(III) complex. Chem Mater 17:5712–5719
99. Du P, Knowles K, Eisenberg R (2008) A homogeneous system for the photogeneration of hydrogen from water based on a platinum(II) terpyridyl acetylide chromophore and a molecular cobalt catalyst. J Am Chem Soc 130:12576–12577
100. Dempsey JL, Winkler JR, Gray HB (2010) Kinetics of electron transfer reactions of H_2-evolving cobalt diglyoxime catalysts. J Am Chem Soc 132:1060–1065
101. Sakai K, Ozawa H (2007) Homogeneous catalysis of platinum (II) complexes in photochemical hydrogen production from water. Coord Chem Rev 251:2753–2766
102. Ozawa H, Yokoyama Y, Haga M-A, Sakai K (2007) Syntheses, characterization, and photo-hydrogen-evolving properties of tris(2,2′-bipyridine)ruthenium(II) derivatives tethered to a cis-Pt(II)Cl_2 unit: insights into the structure–activity relationship. Dalton Trans 1197–1206
103. Lei P, Hedlund M, Lomoth R, Rensmo H, Johansson O, Hammarström L (2008) The role of colloid formation in the photoinduced H_2 production with a Ru^{II}-Pd^{II} supramolecular complex: a study by GC, XPS, and TEM. J Am Chem Soc 130:26–27
104. Du P, Schneider J, Li F, Zhao W, Patel U, Castellano FN, Eisenberg R (2008) Bi- and terpyridyl platinum(II) chloro complexes: molecular catalysts for the photogeneration of hydrogen from water or simply precursors for colloidal platinum? J Am Chem Soc 130:5056–5058
105. Masaoka S, Mukawa Y, Sakai K (2010) Frontier orbital engineering of photo-hydrogen-evolving molecular devices: a clear relationship between the H2-evolving activity and the energy level of the LUMO. Dalton Trans 39:5868–5876
106. Ozawa H, Kobayashi M, Balan B, Masaoka S, Sakai K (2010) Photo-hydrogen-evolving molecular catalysts consisting of polypyridyl ruthenium(II) photosensitizers and platinum(II) catalysts: insights into the reaction mechanism. Chem Asian J 5:1860–1869
107. Tschierlei S, Presselt M, Kuhnt C, Yartsev A, Pascher T, Sundström V, Karnahl M, Schwalbe M, Schäfer B, Rau S, Schmitt M, Dietzek B, Popp J (2009) Photophysics of an intramolecular hydrogen-evolving Ru-Pd photocatalyst. Chem Eur J 15:7678–7688

Hydrothermal Systems, Geochemistry of

DAVID NIEVA, ROSA MARÍA BARRAGÁN,
VÍCTOR ARELLANO
Instituto de Investigaciones Eléctricas, Cuernavac,
Morelos, Mexico

Article Outline

Glossary
Definition of the Subject
Introduction
Geothermometry
Geochemical Modeling of Hydrothermal Fluids
Future Directions
Bibliography

Glossary

Adiabatic process A process undergone by a system during which there is no heat exchange between the system and its surroundings.

Enthalpy Thermodynamic potential (akin to the internal energy) of a system that remains constant when the system undergoes an adiabatic process at constant pressure and composition.

Geothermal reservoir engineering Branch of engineering that applies scientific principles to the problems arising during the extraction of mass and heat from a geothermal resource.

Geothermal resource, Unless otherwise specified, in this article the meaning is restricted to a hydrothermal system selected for exploitation with the purpose of extracting usable heat.

Magma chamber Large accumulation of molten rock found beneath the surface of the Earth.

Definition of the Subject

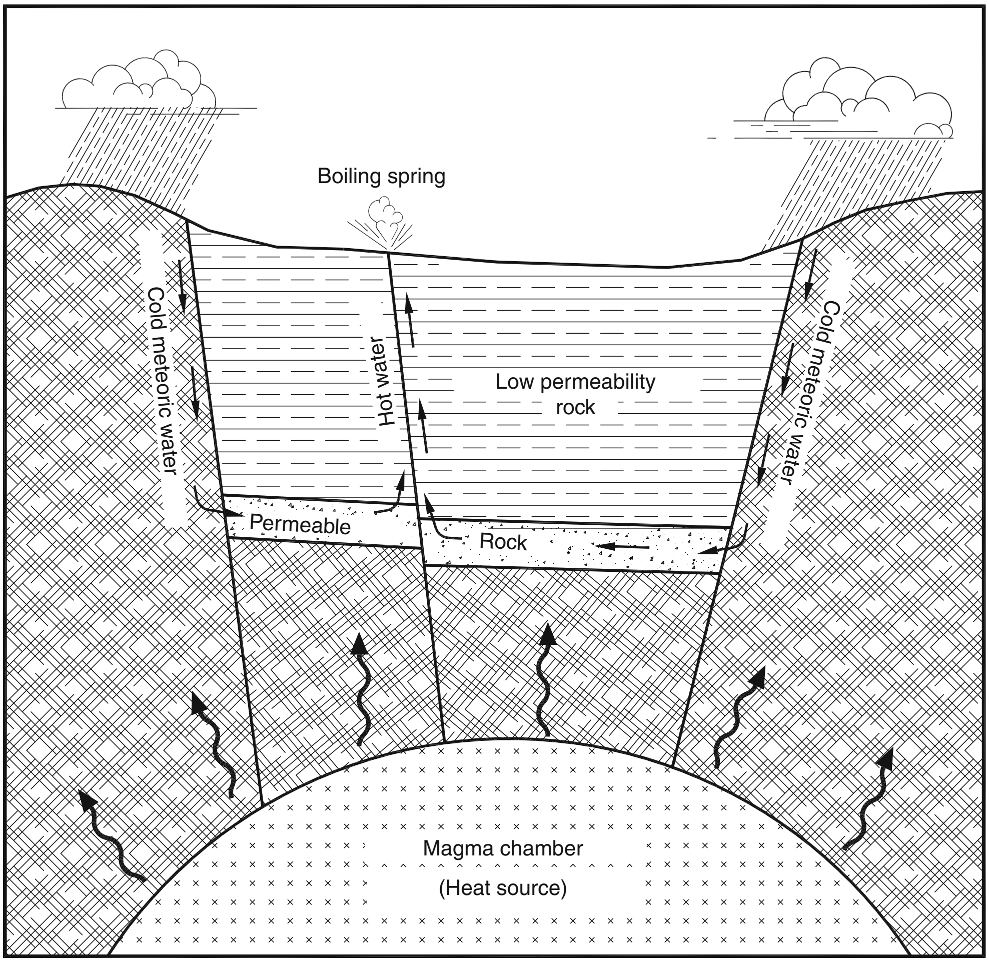
In its most ample definition, geochemistry involves the study of the abundance, distribution, transformation, and transport of the elements, and their isotopes and compounds in the Earth and other planets. In this entry, this extremely ample subject is drastically reduced by two constraints. In the first instance, it is restricted to the discussion of hydrothermal systems, the only type of high-temperature geothermal resource that is exploitable with current conventional technology for the production of electricity and process heat. The second constraint consists of the pragmatic approach to the subject. Indeed, the entry refers to the interplay between science and technology, which has led to the development of techniques based on geochemical considerations that have proven useful for the characterization and exploitation of hydrothermal systems.

Early studies of hydrothermal systems were concerned with the underground flow of heat and water [1–4]. Soon afterward, the systematic study of geochemical features of these systems started [5–7]. Since then, geochemistry has accompanied the development of technology for the exploration and exploitation of geothermal resources. Geochemical techniques provide information on the structure and the distribution of temperature and other physicochemical parameters in a hydrothermal system. When combined with reservoir engineering considerations, geochemistry provides information on processes induced in the geothermal reservoir by the continuous extraction of mass and heat.

Introduction

A hydrothermal system is a volume of rock which is permeated by water or steam and heated to a temperature higher than normal for the particular depth. The source of heat is usually a magma chamber lying a few kilometers underneath; less frequently, the source of heat is a crustal site with an abnormally high concentration of radioisotopes, mainly potassium (^{40}K), thorium (^{232}Th), and uranium (^{235}U and ^{238}U). The origin of the fluid is meteoric, that is, rain water that infiltrates the ground to depths of up to a few kilometers, frequently mixed with fluids and solutes of magmatic origin. The essential features of a hydrothermal system are shown in Fig. 1, which is similar to an early sketch by White [8]. A third feature is the so-called cap rock, which is a layer of less permeable rock overlying the volume of permeated rock (i.e., the "reservoir"). The permeability and degree of fracturing of this cap rock varies from site to site and so do the intensity and the abundance of the surface manifestations of hydrothermal systems (boiling springs, steam vents, hot ponds, geysers).

A more detailed model from a geochemical perspective is that shown in Fig. 2 [9]. Degassing of magma produces a fluid that, on the basis of the analysis of high-temperature gases released from basaltic [10] and andesitic magmas [11] is estimated to consist of water vapor with 20–40% by weight of CO_2, 5% of sulfur (SO_2/H_2S), and 1–2% HCl. As this fluid ascends, mixes with deeply circulating ground water and experiences a reduction in temperature, it forms a highly acidic fluid which interacts with the wall rock in a process called "primary neutralization." The result of this process is postulated to be a virtually complete cation leaching of the rock, forming close to neutral solutions with cation contents similar to the composition of the original rock, and with chloride and sulfate as the main anions. It is postulated further that magnesium and calcium are rapidly redeposited in the form of Mg- and Ca- rich minerals and sulfate is also deposited in the form of pyrite, alunite, or anhydrite. The result of this "primary neutralization" process is the production of the typical brine encountered in geothermal reservoirs, that is, a neutral or close to neutral aqueous solution with chloride as the main anion and sodium and potassium as the main cations. This brine, which contains much of the magmatic CO_2 and some of the sulfur as H_2S [9], may reach the surface mixed with varying proportions of meteoric water in natural hot springs or could be brought to the surface by man-



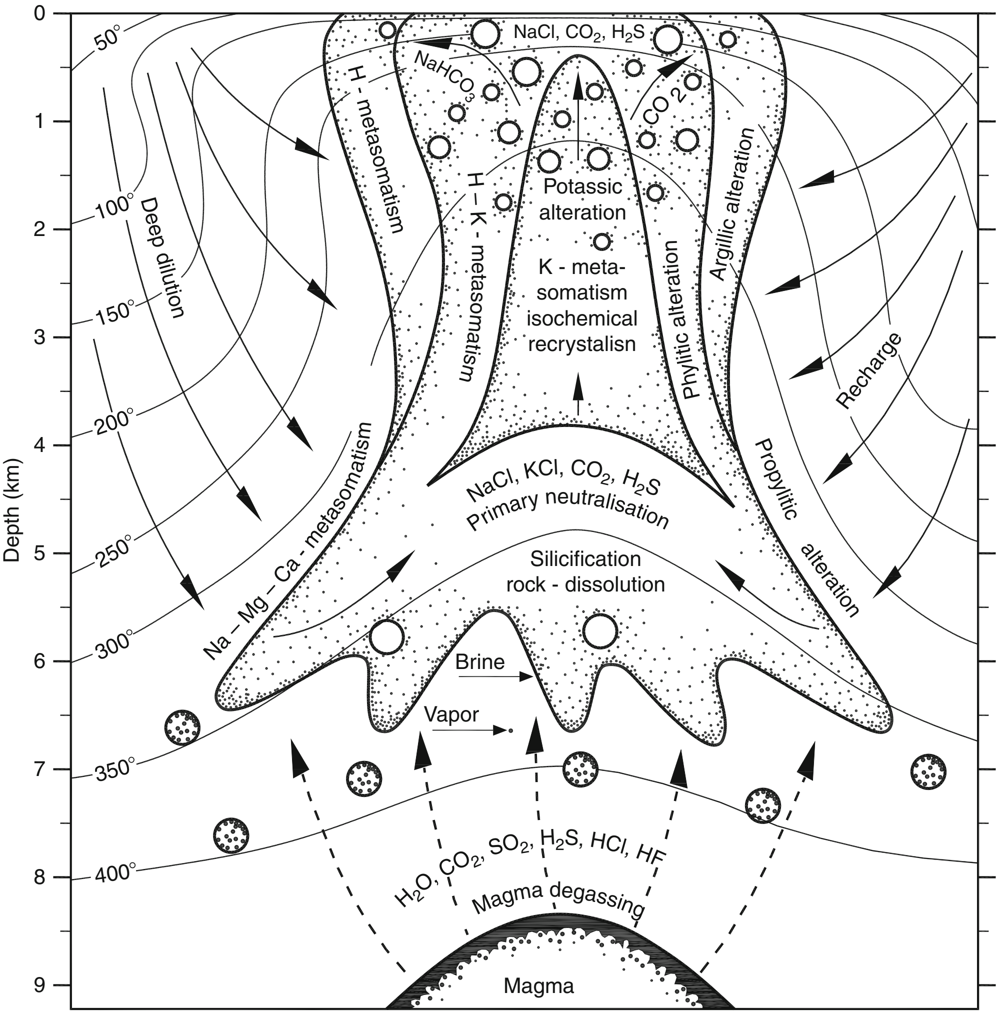
Hydrothermal Systems, Geochemistry of. Figure 1
Schematic model of a hydrothermal system

made wells. The natural upward migration of this NaCl/KCl brine through the main upflow zone (i.e., the "geothermal reservoir") causes some characteristic alteration of the rock which leads to its enrichment in potassium (K-metasomatism).

In some hydrothermal systems, the deep, pre-neutralized liquid phase, finds its way to the surface, mixed with meteoric waters. The oxidation by atmospheric oxygen of the H_2S to produce H_2SO_4 gives rise to surface discharges with very low pH and high concentrations of chloride and sulfate ions; the cation composition of these waters may approach the composition of the original (dissolved) rock [12].

A third type of surface discharge originates from CO_2-rich steam separation from the rising brine, its interaction with the rock at shallower levels to produce calcite, and its mixture with oxygen-rich meteoric waters, which produce surface discharges with bicarbonate as the dominant ion, close to neutral pH, and varying concentrations of sulfate and chloride ions. The mineralogical consequences of this acid-induced alteration (H-metasomatism) of the rock are observable in the periphery of the main upflow zone.

A hydrothermal system is clearly not in equilibrium; there are thermal, mechanical, chemical, and nuclear processes that lead to conductive, convective, and radiative transport of energy, and to mass transport in and out of the system as well as between parts of the system (fluid flow, dissolution and precipitation of material, chemical reactions in liquid, gaseous, and solid phases).



Hydrothermal Systems, Geochemistry of. Figure 2
Schematic cross section of a hydrothermal system with details of fluid genesis and fluid-rock interaction, as drawn by Giggenbach [9]

However, in certain stages of development, and certain parts of the system, where temperature and pressure vary only very slowly, fast chemical reactions adjust to the conditions prevailing at any given time and may be said to be at equilibrium. Many of the useful geochemical inferences that may be made on the prevailing physicochemical conditions in a hydrothermal system depend on these partial equilibrium states, which may be best referred to as *steady states*.

In the early stages of exploration of a hydrothermal system, when only evidence available from the surface is at hand, the aim of a geochemical survey is the generation of a model that would assess the temperature and chemical conditions of the fluid at depth. Naturally, the confidence that may be placed on this assessment increases with the ability of the model to accommodate as much of the available evidence as possible, and to explain the composition of all surface discharges [13–16].

Geothermometry

Perhaps the single most important hydrothermal reservoir parameter that could be estimated with geochemical techniques is the temperature. Indeed, there are algorithms, called geothermometers, which relate the geothermal reservoir temperature to mathematical expressions involving the concentrations of ionic and neutral species in the hydrothermal fluid. In this section, a general discussion of the bases and application of geothermometers is provided.

Cation Geothermometry

Geothermometers based on the ionic composition of hydrothermal fluids explicitly or implicitly refer to mineral–solution reactions of the following type:

$$\begin{aligned}&\text{Mineral assemblage } 1 + \mathrm{aA}^{w+} + \mathrm{bB}^{x+}\\ &\quad = \text{Mineral assemblage } 2 + \mathrm{cC}^{y+} + \mathrm{dD}^{z+}\end{aligned} \tag{1}$$

where A, B, C, and D stand for cations in solution; a, b, c, and d are stoichiometric coefficients, and the charge conservation condition: $aw + bx = cy + dz$ is obeyed.

To a first approximation, the activities of minerals in contact with hydrothermal solutions may be taken as constant and equal to unity. Also, many hydrothermal aqueous solutions are only moderately concentrated, so the activity coefficients of ionic species in solution approach unity. With these simplifications, the equilibrium constant for reaction 1 is given by

$$Keq = \frac{[C^{y+}]^c [D^{z+}]^d}{[A^{w+}]^a [B^{x+}]^b} \tag{2}$$

where the quantities in brackets refer to molal concentrations.

The natural logarithm of the equilibrium constant is related to the reaction changes in standard enthalpy and entropy by the following expression:

$$\ln Keq = -\frac{\Delta \mathrm{H}^0}{\mathrm{RT}} + \frac{\Delta \mathrm{S}^0}{\mathrm{R}} \tag{3}$$

where ΔH^0 and ΔS^0 are the reaction change in standard molal enthalpy and entropy, respectively, T stands for the absolute temperature and R is the gas constant. For many reactions, ΔH^0 and ΔS^0 vary only slightly with temperature and pressure in the range of interest (25–300°C; 1–200 bar), so one would expect to find an approximately linear relationship between the logarithm of the equilibrium constant and the reciprocal of the absolute temperature.

The first indicator of this sort that was identified is the ratio of the concentrations of sodium and potassium ions in the hydrothermal fluid, as it was observed that the ratio K/Na increases with temperature [5, 6]. Mercado [17] generated a model of natural flow in the Cerro Prieto geothermal field (Mexico) based on this indicator.

Several relations of the type:

$$\ln \frac{[\mathrm{K}^+]}{[\mathrm{Na}^+]} = -\mathrm{A}\left(\frac{1}{\mathrm{T}}\right) + \mathrm{B} \tag{4}$$

where A and B are constants, have been fitted to data on fluid chemical composition and measured underground temperature from geothermal wells, oil field brines, and natural hot springs [5, 18–24]. Those relations which were fitted to sufficiently ample data bases from geothermal wells and oil field brines [22–24] approach the expression that would be obtained for Eq. 3 from modern thermodynamic data [25] for the reaction:

$$\text{albite} + \mathrm{K}^+ = \mathrm{K} - \text{feldspar} + \mathrm{Na}^+ \tag{5}$$

Fournier and Truesdell [19] developed a geothermometer based on two hypothetical mineral–solution reactions represented by the single expression:

$$\begin{aligned}&\text{Mineral assemblage } 1 + \mathrm{K}^+\\ &\quad = \text{Mineral assemblage } 2 + (\beta/2)\mathrm{Ca}^{2+} + (1-\beta)\mathrm{Na}^+\end{aligned} \tag{6}$$

where β could take one of two possible values, namely, 1/3 and 4/3. Those authors proposed a graphical method to estimate reservoir temperatures in a plot of the type of Eq. 3 for the above reactions. An analytical expression for this geothermometer, provided by Fournier [26], is as follows:

$$\log_{10} \frac{(\mathrm{Na}^+)^{1-\beta}(\mathrm{Ca}^{2+})^{\frac{\beta}{2}}}{\mathrm{K}^+} = 1647\left(\frac{1}{T}\right) - 2.47 - 2.06\beta \tag{7}$$

where the parentheses on the left-hand term refer to concentrations in mg/kg. The procedure proposed by the authors is to apply Eq. 7, on a first trial, utilizing a value of 1/3 for the parameter β. In case the result obtained is a temperature greater than 100°C, one is to consider it the final value; otherwise, one should utilize the value of 4/3 for β. As pointed out by Giggenbach [9], the fact that a single value for the coefficient of (1/T) is utilized implies that β is to be considered a statistical weighing factor rather than a stoichiometric coefficient in reaction 6.

This Na–K–Ca geothermometer has been applied extensively in the exploration of hydrothermal resources. Several authors have identified systematic errors in its application which seem to be associated with high contents of CO_2 in the hydrothermal fluid [9, 27, 28], and a modified expression has been proposed [27]. In addition, Fournier and Potter [29] have proposed a correction term to be used when this geothermometer is applied to Mg-rich thermal waters.

Geothermometers based on the concentration of the four main cations (Na^+, K^+, Ca^{2+}, Mg^{2+}) have been developed. Nieva and Nieva [24] developed a procedure that leads to the utilization of one of four possible equations with a given sample, depending on its chemical characteristics. This geothermometer, called the Cationic Composition Geothermometer (CCG), is designed to ensure that the equation:

$$\log_{10}\frac{[K^+]}{[Na^+]} = -1178\left(\frac{1}{T}\right) + 1.239 \tag{8}$$

similar to the expression developed by Fournier [22], is applied only to waters showing preliminary signs of having a high-temperature origin, thus adopting a conservative approach which avoids overestimation of temperatures at depth. Application of this geothermometer to a hot-spring NaCl/KCl brine with close to neutral pH is safe and the resulting estimated temperature should be considered a lower bound for the reservoir temperature. In Eq. 8 brackets indicate molal concentrations.

Giggenbach [9] also developed a procedure to estimate temperatures at depth based on the consideration of chemical equilibria involving the four main cations. This author concluded that the mineral assembly muscovite, clinochlore, silica, albite, and K-feldspar is at chemical equilibrium in the deep parts of the main upflow zone of the hydrothermal system according to the following reactions:

$$\begin{aligned}&0.8\,\text{muscovite} + 0.2\,\text{clinochlore} + 5.4\,\text{silica} + 2\,Na^+\\ &\quad = 2\,\text{albite} + 0.8\,K-\text{feldspar} + 1.6\,\text{water} + Mg^{2+}\end{aligned} \tag{9}$$

and

$$\begin{aligned}&0.8\,\text{muscovite} + 0.2\,\text{clinochlore} + 5.4\,\text{silica} + 2\,K^+\\ &\quad = 2.8\,K-\text{feldspar} + 1.6\,\text{water} + Mg^{2+}\end{aligned} \tag{10}$$

The above reaction system includes, of course, reaction 5. From these considerations, and taking as a reference modern thermodynamic data, Giggenbach [9] proposed the use of a K-Na geothermometer and a K-Mg geothermometer, as given respectively by the following expressions:

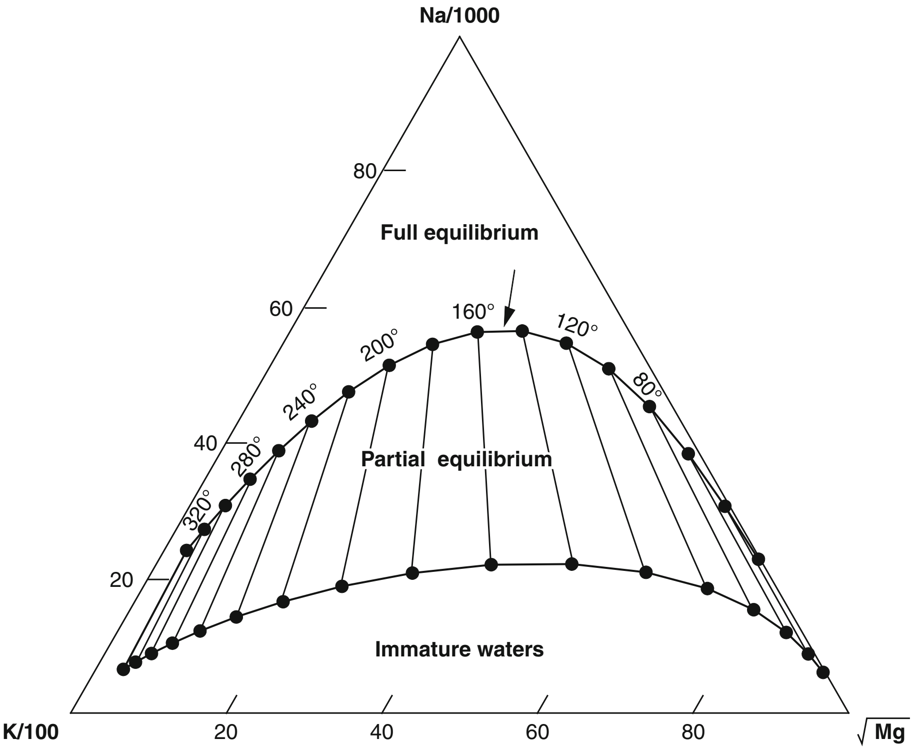
$$\log_{10}\frac{[K^+]}{[Na^+]} = -1390\left(\frac{1}{T}\right) + 1.75 \tag{11}$$

and

$$\log_{10}\frac{[K^+]^2}{[Mg^{2+}]} = -4410\left(\frac{1}{T}\right) + 14 \tag{12}$$

In Eqs. 11 and 12 concentrations are in mg/kg. Reaction 5 seems to re-equilibrate very slowly as the fluid ascends and encounters strata with lower temperatures. Thus, as mentioned above, K-Na geothermometers applied to waters from hot springs provide a reasonable lower bound for reservoir temperature at depth. On the other hand, the proportion of magnesium and potassium ions readjusts more rapidly and the application of Eq. 12 yields estimates of temperature in shallower parts of the hydrothermal system [9].

Just as in the case of the development of the CCG geothermometer [24], Giggenbach [9] faced the need to develop an indication, based on the concentration of Mg^{2+}, to determine if thermal water had been at equilibrium at depth. This author proposed the use of the triangular diagram shown in Fig. 3, based on the parameters $Na^+/1{,}000$, $K^+/100$, $(Mg^{2+})^{1/2}$. In this diagram, data points of waters such as those produced by geothermal wells, which have a cationic composition compatible with chemical equilibrium of reactions 9 and 10, fall on the "fully equilibrated waters" curve. In contrast, data points of water with composition compatible with full dissolution of crustal rocks fall in the lower right-hand corner. The region in the



Hydrothermal Systems, Geochemistry of. Figure 3
Triangular diagram for estimation of equilibrium reservoir temperatures by simultaneous application of geothermometric expressions from Eqs. 11 and 12. The "full equilibrium" curve represents the loci of intersections of isotherms drawn by use of Eqs. 11 and 12. Dot-joining lines are isotherms drawn by use of Eq. 12. Figure constructed with information provided by Giggenbach [9]

diagram belonging to "partially-equilibrated waters" is bounded on the top by the "fully equilibrated waters" curve and on the bottom by a rather arbitrary curve. This lower curve is drawn by plotting data as for the "fully equilibrated waters" but first dividing the concentration of sodium by 4.57. Giggenbach [9] recommends the use of the K-Mg geothermometer for samples that fall in the partially-equilibrated waters area of the diagram, but to exercise caution when they fall near the lower curve.

Table 1 illustrates the application of the CCG [24] and the Na-K-Mg [9] geothermometers, as applied to samples from geothermal wells and hot springs from a number of hydrothermal systems. For most wells, the CCG yields temperatures quite close to the measured values. When applied to hot springs, the CCG yields temperatures that, in the context of a survey of a geothermal area, could be safely taken as a lower bound for the reservoir temperature. The Na-K-Mg geothermometer yields temperatures close to the measured value but in some cases, it overestimates the temperature of the reservoir. For this reason, some authors [30, 31] prefer to construct the triangular diagram of Giggenbach [9] using Eq. 8 of the CCG instead of Eq. 11 of the Na-K-Mg geothermometer. Cationic geothermometers fail when applied to ocean water, yielding high-temperature values possibly reflecting the interaction of ocean water with basaltic rock in mid-ocean ridges [9]. In cases of hydrothermal systems with an important intrusion of ocean water, care must be taken in the application of cationic geothermometers; their application should be limited to brines produced by deep wells which presumably have been subject to interaction with reservoir rocks at high temperatures for long periods of time.

Hydrothermal Systems, Geochemistry of. Table 1 Application of the CCG [24] and the Na-K-Mg [9] geothermometers to geothermal wells and associated hot springs. Concentrations are in mg/kg and temperatures are in Celsius. Data is from [9]

Origin	Reservoir Temperature	pH	Na	K	Mg	Ca	t CCG	t K-Na	t K-Mg
Geoth. well NG-9, Ngawha, New Zealand	230	7.7	893	79	0.110	3	194	223	204
Jubilee pool, New Zealand	230	7.2	842	72	1.600	9	191	220	147
Geoth. well WK66, Wairakei, New Zealand	240	8.5	995	142	0.040	17	236	262	258
Champagne Pool, New Zealand	240	8.0	1,070	102	0.400	26	200	229	187
Geoth. well BR-11, Broadlands, New Zealand	260	7.4	675	130	0.011	1	266	291	291
Ohaki Pool, New Zealand	260	7.1	860	82	0.100	3	200	229	208
Geoth. well 10, Miravalles, Costa Rica	250	7.8	1,750	216	0.110	59	222	250	254
Sal. Bagaces, Costa Rica	250	8.6	2,063	85	6.600	33	140	170	129
Geoth. well ZQ-3, Zunil, Guatemala	295	8.1	933	231	0.012	15	294	317	327
Zunil Spring	295	7.0	384	32	39	17	189	218	77
Geoth. well from Cerro Prieto, Mexico	280	7.3	5,600	1,260	0.270	333	283	307	337
Spring N29, Cerro Prieto, Mexico	280	7.6	5,120	664	4.600	357	227	254	216
Geoth. well from Tongonan, Philippines	330	7.0	3,580	1,090	0.200	128	320	340	337
Banati Spring, Philippines	330	8.3	1,990	210	0.400	86	208	237	219
Salton Sea well, USA	330	5.2	38,400	13,400	10	22,010	338	357	381
Seawater	4	7.8	10,760	390	1,290	410	132	162	97

Most cation geothermometers involve only the major cations (Na^+, K^+, Mg^{2+}, Ca^{2+}) but some are based on lithium equilibrium between the aqueous solution and some mineral assembly [32, 33]. Lithium–magnesium geothermometers yield values close to the measure reservoir temperature when applied to water produced by geothermal and oil wells; however, when applied to hot spring water, the calculated temperatures are frequently too low, presumably due to the re-equilibration at lower temperatures of reactions involving magnesium [26].

Silica Geothermometry

The concentration of silica in hydrothermal fluids is determined by the solubility of quartz [34]. The fact that the solubility of quartz shows strong temperature dependence [34–36] provides the basis for the silica geothermometer [37, 38], which has been applied extensively for the exploration of geothermal resources.

Two basic assumptions underlie the application of the silica geothermometer. The first assumption is that the solubility of quartz in moderately saline hydrothermal fluids is similar enough to its solubility in pure water, so that the latter could be taken as reference. This assumption is quite plausible as long as the salinity is below 2% by weight and temperatures are below 300°C, which is the case for many hydrothermal systems; a correction must be made when salinity or temperature are higher [26].

The second assumption is that the hydrothermal fluid does not gain or lose silica after it has reached equilibrium with quartz in the reservoir. This assumption does not hold in cases when the hydrothermal fluid undergoes a relatively slow rise to the surface. This is the reason why the silica geothermometer applied to hot spring waters does not yield temperatures higher than about 230°C; the reactions of quartz dissolution or precipitation are very fast above this temperature [39, 40]. However, when applied to hydrothermal fluids produced by lined wells that deliver the fluid to the surface much more rapidly and without interaction with the rock, the silica geothermometer is quite useful to estimate the reservoir temperature.

In addition to the above assumptions, when applying the silica geothermometer to hot spring waters, one must make some conjecture as to the processes affecting the sample on its way to the surface. When reaching the point of discharge, the hydrothermal liquid must have reduced its temperature to some value at or below the boiling point. There are basically three mechanisms by which this may be accomplished: mixing with colder waters, loss of heat to the rock by conduction, and adiabatic separation of steam; in most cases, at least two of these mechanisms intervene. When modeling these processes, it is much more convenient to deal with enthalpy (i.e., heat content) rather than temperature because the former is a conservative quantity. Thus, when the hydrothermal liquid loses heat to the rock by conduction, the heat transferred is subtracted from the enthalpy of the liquid. When the hydrothermal liquid mixes with a colder water, the enthalpy (and silica) content of the mixture is simply the sum of the quantities contributed by each of the mixing components. Likewise, when a hydrothermal fluid undergoes adiabatic steam separation at a given temperature, the enthalpy distributes itself between the resulting liquid and steam phases according to the specific enthalpy (i.e., the heat content per unit mass) of each phase:

$$H_{hf} = (1-f)\bar{H}_{t,l} + f\bar{H}_{t,s} \qquad (13)$$

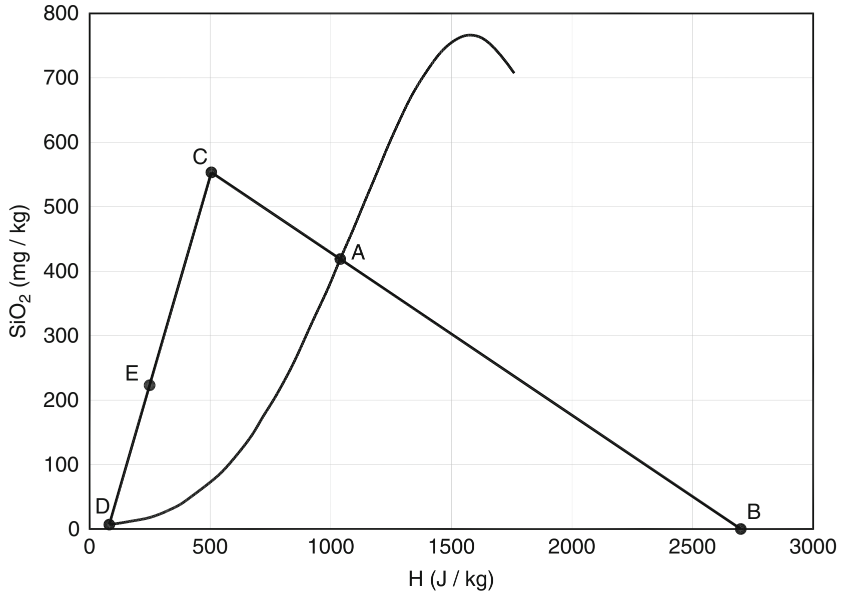
where f is the mass fraction of steam, H_{hf} stands for the specific enthalpy of the hydrothermal liquid before undergoing steam separation, and $\bar{H}_{t,l}$ and $\bar{H}_{t,s}$ refer to the specific enthalpies at temperature t of saturated liquid and steam, respectively. Specific enthalpies of saturated steam and liquid are usually taken as those for pure H_2O and could be obtained from steam tables. The silica content of the hydrothermal liquid also distributes itself between the separated phases:

$$[\mathrm{SiO_2}]_{hf} = (1-f)[\mathrm{SiO_2}]_l + f[\mathrm{SiO_2}]_s \qquad (14)$$

When the steam separation occurs at temperatures of 260°C or less, the silica concentration in the separated steam could be approximated safely as zero.

These processes could be modeled graphically in a Cartesian plot of silica concentration vs. enthalpy. Fournier and Potter [37] provide a plot of the equilibrium silica concentration in H_2O at the vapor pressure of the solution, as a function of enthalpy, covering the range of liquid water at 20°C to temperatures close to the critical point; Fig. 4, which reproduces a portion of that plot, was constructed with information provided in the cited reference. In such a plot, a point representing the hydrothermal liquid would fall on the quartz solubility curve. A point representing the mixture of the hydrothermal liquid with a cold water component would fall on the tie line joining the points which represent the two mixing components. In order to represent a process of adiabatic steam separation, one would locate the point representing the separated steam, with zero silica concentration and the specific enthalpy of the steam; then one would draw a line joining this point with the point representing the original hydrothermal liquid; by extending this line to the value of the specific enthalpy of the separated (saturated) liquid, one would find the resulting silica concentration and thus the position of the point representing this liquid. As an example, the following process is depicted graphically in Fig. 4: A hydrothermal liquid at 240°C undergoes adiabatic cooling by expansion with single-stage steam separation at 120°C. The separated liquid then mixes (w/w) with 60% meteoric water at 15°C, containing 5 mg/kg of silica, giving rise to a thermal water at about 57°C and a silica concentration of about 224 mg/kg.

As mentioned above, in spite of the potential complexity of the processes affecting the silica content of hot spring fluids the silica geothermometer has been applied extensively for the exploration of geothermal resources. When applied to hydrothermal liquid produced by geothermal wells, the situation is particularly simple. The liquid is brought to the surface rapidly, with negligible loss of heat to the walls and without mixing with other fluids; the only mechanism that intervenes is the adiabatic separation of steam. For these reasons, the silica geothermometer applied to well fluids is a powerful tool for the estimation of the reservoir temperature.



Hydrothermal Systems, Geochemistry of. Figure 4
Graph of equilibrium silica concentration in H2O at the vapor pressure of the solution vs. specific enthalpy. A process of adiabatic cooling of a hydrothermal liquid (point A), followed by mixing with a local meteoric water component (point D), is depicted graphically (see main text). Points (A, B, C, D, E) and compositions (specific enthalpy in J/kg, dissolved silica in mg/kg) are as follows: Original hydrothermal liquid: A (1,037, 418); steam separated adiabatically at 120°C: B (2,706, 0); liquid separated adiabatically at 120°C: C (503.7, 552); cold meteoric water: D (62.99, 5); resulting thermal water: E (239.3, 224)

Gas Geothermometry

Reactions that involve gaseous substances that are present and measurable in the hydrothermal fluid may provide the basis for gas geothermometers. These reactions may or may not involve solid (mineral) phases. Examples of reactions which take place in the hydrothermal fluid and do not involve a mineral phase are the so-called Fischer–Tropsch reaction:

$$CH_4 + 2H_2O = CO_2 + 4H_2 \tag{15}$$

and the conversion of nitrogen to ammonia:

$$2NH_3 = N_2 + 3H_2 \tag{16}$$

An example of a gas reaction which involves mineral phases is the oxidation of pyrite to produce magnetite:

$$3FeS_2 + 2H_2 + 4H_2O = Fe_3O_4 + 6H_2S \tag{17}$$

The collection and handling of samples for the analysis of gaseous species in hydrothermal fluids is, in general, more complicated than is the case for species completely dissolved in the liquid phase [41]. Furthermore, the relative proportions of the gaseous species sampled in the hydrothermal fluid may have been altered by a number of untraceable processes that affect the fluid on its transit from the reservoir to the surface. All these circumstances introduce uncertainty in the interpretation of these analyses, as will be discussed below.

The equilibrium proportions of chemical species that participate in a reaction at a given temperature in, say, the steam phase, are known from thermochemical data. It is safe usually to assume that there exists phase equilibrium between liquid and steam phases, so from the knowledge of liquid–steam partition coefficients, one may know also the equilibrium proportions of species in the liquid phase.

The most favorable situation for the interpretation of the gaseous composition of the hydrothermal fluid is one in which the undisturbed liquid phase is sampled directly from the reservoir. This situation is approximated closely when one knows that a geothermal well is fed only by liquid phase. In such situation, one may determine the temperature of equilibration of the chemical reaction (i.e., the reservoir temperature) from the analyses of the fluid and the temperature dependence of the equilibrium constant (Eq. 3).

In most cases, however, a geothermal well receives a mixture of liquid and steam phases from the reservoir. This leads to a situation that is considerably more complex [42] because the mobilities of liquid and steam phases in the porous medium are different; thus, the proportions of liquid and steam phases reaching the well will be somewhat different than the proportions in the undisturbed reservoir. In spite of these drawbacks, in these situations the gas composition of the produced fluid is utilized to estimate the fractions of reservoir steam and liquid that make up the mixture feeding the well [43–45]. The fraction of reservoir steam that results from these calculations, which is referred to as the *reservoir steam fraction* and frequently represented by the letter *y*, approximate as it may be, is a useful reservoir parameter, particularly when employed to detect variations in the conditions of the reservoir resulting from exploitation.

The reservoir temperature, which is a necessary piece of information for the calculations described in the last paragraph, is measured directly, or estimated by the use of other geothermometers. D'Amore and Truesdell [46] developed a graphical method to calculate *both* the reservoir temperature and the reservoir steam fraction by considering simultaneous equilibrium in two reactions in the gas phase.

Some of the reactions that serve as a basis for the calculation of the reservoir steam fraction, namely, the Fischer–Tropsch reaction (Eq. 15) and the conversion of nitrogen to ammonia (Eq. 16), have been reported NOT to be at equilibrium in hydrothermal systems with reservoir temperatures below 250°C [47, 48], so these calculations must be interpreted with caution for such systems.

If the interpretation of the gas composition of fluids produced by geothermal wells is complicated, the case of natural steam vents is even more so. Indeed, there are several processes that may occur during the rise of the hydrothermal fluid to the surface that may affect its composition. For example, the hydrothermal fluid may lose steam at different stages (and different pressures and temperatures) during its rise; the rising steam–liquid mixture may suffer partial condensation at different stages; also, partial re-equilibration of chemical reactions may occur in transit to the surface. Nevertheless, Arnorsson and Gunnlaugsson [47] developed a number of geothermometers based on the concentration of gaseous species, and applied them to steam vents in Icelandic geothermal fields. For some of these sites, the calculated temperatures compare surprisingly well with temperatures measured in nearby wells, but this may be due to particularly favorable hydrological conditions.

Giggenbach [49] proposes a geothermometer based on the ratio of molar concentrations of argon and hydrogen in the hydrothermal fluid, which is designed to be relatively insensitive to processes of steam separation or partial condensation occurring as the hydrothermal fluid ascends toward the natural steam vents. Since both argon and hydrogen are very insoluble in water, and would separate almost quantitatively from the liquid phase upon separation of even a small mass fraction of steam, they are likely to keep their relative proportions unchanged during the rise of the hydrothermal steam to the surface. This geothermometer is based on the following main assumptions. (a) In liquid-dominated systems, the chemical equilibration of species, including volatile components, occurs in the liquid phase. (b) The redox conditions in a fully equilibrated hydrothermal system are quite invariant and keep the ratio of the fugacities of hydrogen to water at a value of approximately $10^{-2.8}$ at temperatures up to 1,200°C. (c) The argon concentration in the reservoir liquid corresponds to that of air-saturated groundwater. With these assumptions, Giggenbach [49] arrives to the expression:

$$t(^{o}C) = 71\left(\log_{10}\frac{X_{H2}}{X_{Ar}} + 2.5\right) \quad (18)$$

where X_i stands for the molar fraction of species i in the sampled fluid. This geothermometer was tested with appreciable success in samples from New Zealand and elsewhere. One serious drawback with this geothermometer is the fact that even the smallest contamination of the sample with air may lead to very erroneous results. Another problem is the possibility that, during its rise to the surface, the fluid may have lost argon and hydrogen by steam separation, followed by recuperation of hydrogen via chemical reequilibration, thus leading to an erroneous high ratio of hydrogen to argon, and thus to an overestimation of the reservoir temperature.

It is fair to say that for the exploration of a geothermal site, gas geothermometers should be interpreted with extreme care, and only in conjunction with other techniques such as the cationic or silica geothermometers. This may not be possible in sites where most or all the surface manifestations consist of steam vents and steam-heated waters, in which case the gas geothermometer results should be taken as a rough reference.

Solid Solution Geothermometry

Some minerals tend to form solid solutions through substitution of ions at specific sites in the crystal structure, or by inclusion of ions at interstitial spaces. These ion substitutions or inclusions do not change the crystal structure except for rather minor dimensional distortions. The extent of ion replacement depends on various physical and chemical factors, such as the composition of the surrounding medium, the temperature, and the pressure. In principle, this property of minerals offers the opportunity to identify solid-solution geothermometers when the dominating factor that controls the composition of the solid solution is the temperature.

When drilling a geothermal well in a new area, where there are no nearby wells that provide an indication of the temperature vs. depth profile, a relatively accurate temperature estimation that one could obtain, say from the study of drill cuttings, would be extremely useful. This is so because the information would help decide when to stop drilling and complete construction of the well. At present, engineers need to run costly and time-consuming temperature logs in order to obtain this information. This is an important motivation to develop solid-solution geothermometers.

A geothermometer based on the composition of chlorites has been developed [50–52]. Indeed, linear correlations have been observed of increasing crystallization temperature with increasing aluminum substitution for silicon in the tetrahedral site and with decreasing octahedral site vacancy. In spite of the successful substitution of the original automatic electron microprobe technique with the use of X-Ray Diffractometry [53], the process of sample isolation, analysis, and interpretation is much too complex and time consuming for this geothermometer to be useful in the context described in the preceding paragraph. Nevertheless the geothermometer has been quite useful for the determination of paleotemperatures in a variety of studies.

Geochemical Modeling of Hydrothermal Fluids

In the exploratory stage of development of a hydrothermal resource, the main purpose of geochemical modeling is to determine the temperature and composition of the reservoir fluid, based on the available evidence from surface discharges. This is done by tracing back the processes occurring during the rise to the surface of the hydrothermal fluid. As mentioned above, the reliability of the resulting model will depend on its ability to account for all available evidence, that is, its ability to explain through simple but reasonable processes the origin of all surface discharges, as originating from deep reservoir fluids.

Our ability to model the rise of the hydrothermal fluid depends on our knowledge of the behavior of certain elements and compounds, generally called tracers, upon occurrence of processes of steam–liquid separation and mixing of different fluids.

Stable Isotopes of Hydrogen and Oxygen

Two of those tracers are the stable isotopes of hydrogen and oxygen, deuterium (^2H) and oxygen-18, present in natural waters. Indeed, the species $H_2{}^{18}O$ is present in ocean water in concentrations of the order of 2,000 parts per million. Also, the species HDO, where D stands for deuterium, is present in concentrations of the order of 300 parts per million.

The hydrogen and oxygen isotopic composition of water is measured with a mass spectrometer that determines the ratio of the rare to the common isotopic species (i.e., the ratios HDO/H_2O and $H_2{}^{18}O/H_2O$). Rather than expressing these ratios in absolute terms, they are conveniently expressed in the delta notation (δ) as a per thousand (‰) difference with respect to a water standard.

Thus, the sample's deuterium composition, δD, is calculated as follows:

$$\delta D = \frac{R_{\text{sample}} - R_{\text{standard}}}{R_{\text{standard}}} \times 1000 \qquad (19)$$

where R stands for the ratio HDO/H_2O. The oxygen-18 composition, $\delta^{18}O$, is calculated in a similar manner.

The standard that has been adopted almost universally is Standard Mean Ocean Water (SMOW), a water sample with oxygen and hydrogen isotope ratios close to the mean isotope ratios of ocean water [54].

When water separates in two phases (by condensation of steam or evaporation of liquid water), its isotopic composition undergoes *fractionation*, that is, the isotopic composition of the resulting phases is different than the isotopic composition of the original phase. At atmospheric temperatures, the resulting liquid phase will be enriched in the heavy isotopes, deuterium and ^{18}O, and the resulting steam phase will be depleted of them.

As part of the hydrological cycle, water from the ocean continuously evaporates. As the resulting mass of water vapor moves into the continental area, it condensates and looses part of its mass as precipitation (rain). In every precipitation event, the remaining atmospheric steam is depleted further of the heavy isotopes. The process continues as the mass of atmospheric steam moves further into the continental area, giving rise to a distribution of isotopic composition of surface waters, also referred to as "meteoric" waters. This isotopic composition of meteoric water may be seen as a natural label. It was first pointed out by Craig [55] that the isotopic compositions of meteoric waters, which had not undergone appreciable evaporation, were such that the deuterium and oxygen-18 concentrations were linearly related by the following expression:

$$\delta D = 8\delta^{18}O + 10 \qquad (20)$$

After Craig's original observation and with more data, Taylor [56] observed a better fit to the expression:

$$\delta D = 8\delta^{18}O + 5 \qquad (21)$$

This linear expression is known as the Global Meteoric Water Line (GMWL), and serves as a frame of reference not only for the composition of "fresh" meteoric waters, but also for the isotopic composition of surface waters that have undergone other processes, such as evaporation or isotope exchange with rocks.

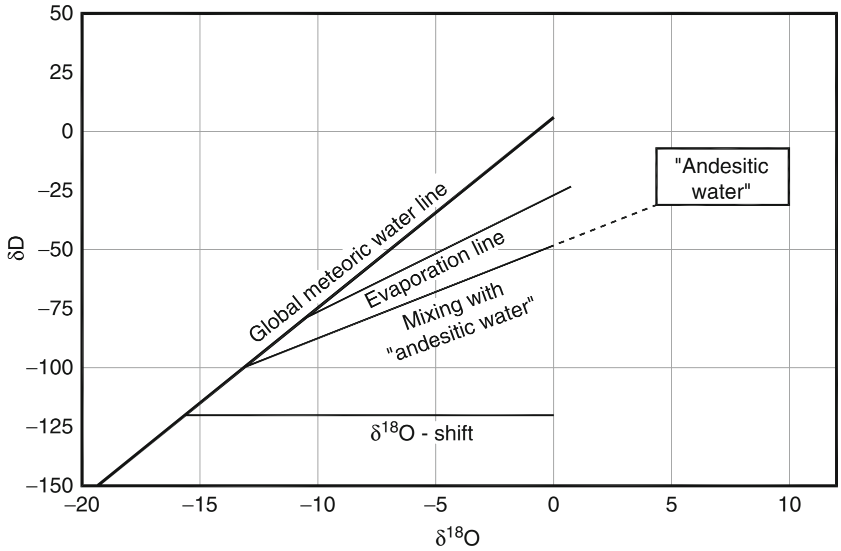
There are some natural processes that may alter the isotopic composition of meteoric waters accumulated in the surface or infiltrated in the ground. One of these processes is evaporation. The isotopic composition of a series of surface waters that have originated from the same meteoric water, but which have undergone evaporation to different degrees, would deviate from the GMWL usually along a line of the form

$$\delta D = y\delta^{18}O + x \qquad (22)$$

where the slope y would have a value between 4 and 6. The residual waters after evaporation would have higher concentrations of the heavy isotopes and therefore would deviate "to the right" of the GMWL in a δD vs. $\delta^{18}O$ Cartesian plot (see Fig. 5).

Craig [57] observed that thermal waters had a deuterium concentration similar to that of local meteoric waters, but had larger concentrations of oxygen-18, a phenomenon that has been termed the "^{18}O shift." This was ascribed to isotope exchange between water and rock, in such a manner that the comparatively small degree of hydrogen exchange leaves the deuterium content of the water virtually unchanged, but its oxygen-18 content is substantially increased. These observations led to the concept that the main component of thermal waters is, by far, meteoric water that has infiltrated the ground to great depth.

Craig made his observations on what was at the time a comprehensive set of data from known thermal sites. However, as data from more thermal areas has accumulated, it has provided grounds to postulate the existence of a deep water component from partial magmatic origin, called "andesitic water," which in hydrothermal systems found in convective plate boundaries mixes with local meteoric water to form the reservoir fluid tapped by geothermal wells [58, 59] (See Fig. 5). According to this alternative proposition, most of the isotopic exchange with the rock occurs at depths greater than those which are normally reached by drilling, presumably in the site where the "primary neutralization" of magma fluids occurs (see "Introduction"), and the result of this neutralization



Hydrothermal Systems, Geochemistry of. Figure 5

δD vs. $\delta^{18}O$ diagram showing the position of meteoric waters (Global Meteoric Water Line) and the positions of meteoric waters subjected to evaporation, oxygen-exchange with the rock, and mixing with a deep magmatic component (andesitic water)

is the formation of andesitic water. At the very high temperatures at which this process occurs (close to magma temperatures), it is expected that the proportion of deuterium (and of oxygen-18) in the andesitic water be completely determined by the isotopic composition of the rock or magma phases. It has been suggested that this composition should fall in the ranges of +5‰ to +10‰ for $\delta^{18}O$ and −30‰ to −10‰ for δD [59–61]. It is further postulated that part of this andesitic water would ascend and mix with more meteoric water to form the hydrothermal or reservoir fluid which is reached by drilling in geothermal fields.

As explained in section on "Introduction," in the natural state of a hydrothermal system, the reservoir fluid ascends to the surface through a variety of processes and gives rise to thermal springs, steam vents, and other features typical of thermal areas. In the absence of drilled wells, these are the only sources of information on the chemical and isotopic characteristics of the reservoir fluid. It is the task of geochemistry to infer the temperature and composition of the reservoir fluid from the available evidence.

As explained in relation to the application of the silica geothermometer (Section on "Silica Geothermometer"), with the consideration of parameters that are conserved in the fluid (silica, enthalpy, chloride ion), it is possible to model the processes affecting the rising fluid on its transit to the surface. Under the reasonable hypothesis that no further isotopic exchange occurs between the fluid and the rock during this transit, the deuterium and oxygen-18 contents of the fluid constitute two more conservative parameters.

When liquid water boils, the resulting liquid and steam phases will in general not have identical isotopic compositions. The parameter that describes the equilibrium distribution of isotopes between the two resulting phases (i.e., the isotopic *fractionation*) is the fractionation factor, α. For a given isotope, the fractionation factor is given by:

$$\alpha = \frac{R_{\text{liquid}}}{R_{\text{steam}}} \tag{23}$$

where R stands for the ratio HDO/H_2O in the case of deuterium, and for the ratio $H_2{}^{18}O/H_2O$ in the case of oxygen-18. For these two isotopes, Truesdell et al. [14] have given a compendium of values of α at temperatures from 0°C to 374°C. The fractionation factor is related to the isotopic compositions of the resulting liquid and steam phases by the following expression:

$$\alpha = \frac{1000 + \delta_{\text{liquid}}}{1000 + \delta_{\text{steam}}} \tag{24}$$

On its rise to the surface, the reservoir liquid may cool by adiabatic boiling, with the steam and liquid remaining in physical contact (and isotopic equilibrium) until the two phases are physically separated at a given temperature. This process shall be called *single-step adiabatic steam separation*. The term adiabatic means that no heat is transferred from or to any surrounding solid (rock) phases. The conservation of enthalpy demands that the mass fraction, y, of generated steam be given by:

$$y = \frac{H_{\text{ol}} - H_{\text{sl}}}{H_{\text{ss}} - H_{\text{sl}}} \tag{25}$$

where H_{ol} represents the specific enthalpy of the original liquid, and H_{sl} and H_{ss} represent the specific enthalpy of liquid water and steam at the separation temperature, respectively.

The conservation of mass for deuterium (or oxygen-18) leads to the following expressions relating the isotopic compositions of the original liquid and the resulting liquid and steam phases:

$$\delta_{\text{ol}} + 1000 = (\alpha + y(1 - \alpha))(\delta_{\text{ss}} + 1000) \tag{26}$$

and

$$\delta_{\text{sl}} + 1000 = \alpha(\delta_{\text{ss}} + 1000) \tag{27}$$

where α represents the value of the fractionation factor at the temperature of separation.

In a real situation, the ascending fluid may suffer more than one event of steam–liquid separation at decreasing temperatures. In fact, one could conceive of a process in which the generated steam separates immediately and continuously from the remaining liquid as soon as it is formed. This process shall be termed *continuous adiabatic steam separation* and, together with the *single-step adiabatic steam separation* described above, constitute the two extreme possible mechanisms of adiabatic boiling of the reservoir fluid on its way to the surface. Both mechanisms lead to an increase in deuterium and oxygen-18 contents of the boiling liquid, but the extent of enrichment is different from one case to the other. An intermediate mechanism that may be called *multiple-stage steam separation* leads, correspondingly, to intermediate enrichments.

The rising reservoir fluid may cool by adiabatic boiling or by mixing with cooler meteoric waters. It may also boil and then mix with meteoric waters, or mix first and then boil, depending on the particular hydrological situation. Truesdell et al. [14] were able to explain the variety of isotopic and chemical compositions of thermal spring waters in Yellowstone National Park, United States, as resulting from a single reservoir liquid at 360°C and particular deuterium and chloride ion contents.

Other Useful Tracers

Other useful tracers are atomic, molecular, or ionic species that exhibit very low chemical reactivity under the conditions prevailing in geothermal reservoirs and upper strata of hydrothermal systems. Some of these tracers partition very favorably (almost exclusively) in the liquid phase, such as chloride ion; some partition very favorably in the steam phase, such as the noble gases, and some distribute themselves in the two phases, but in a manner that can be calculated quantitatively from known partition coefficients (such as the cases of deuterium and oxygen-18). The particular case of silica was discussed in section on “Silica Geothermometer.”

When a relatively high flow of the hydrothermal fluid has passed through some natural conduit to the surface, or through a geothermal well, for a sufficiently long time, the radial distribution of temperature around that conduit or well has reached a steady state. Under these conditions, the transfer of heat between the fluid and the rock could be considered negligible, and the enthalpy of the fluid remains constant (see the appendix of reference [14], for a quantitative discussion on this subject). Under these circumstances the enthalpy is another useful “tracer,” or conserved parameter, that could be quite useful in modeling the processes affecting the fluid on its rise to the surface [62]. In preceding paragraphs, when referring to adiabatic processes, we have implicitly assumed conditions that support the conservation of the enthalpy of the fluid.

The combined use of tracers and conserved parameters has made possible the determination of conditions at depth in hydrothermal systems [13–16, 63].

Future Directions

There are, of course, many scientific questions remaining to be answered regarding the geochemistry of hydrothermal systems, among them the details of the contribution of magmatic fluids to the formation of the hydrothermal brine. From a technological perspective, however, one would expect more future activity in the interplay between geochemistry and reservoir engineering in studies regarding the sustainability of hydrothermal systems under exploitation.

The next type of geothermal resources to be exploited for electric power generation on a large scale will probably be the so-called Enhanced (or Engineered) Geothermal Systems (EGS). These consist of volumes of hot, impermeable rock, located at reachable depths, which contain substantial amounts of heat. Their development as an exploitable reservoir consists of creating a network of fractures, so as to allow the injection of a working fluid (water) at some point and its extraction through wells located at other points, in such a way that the working fluid extracts the heat from the rock. Important practical issues will arise as a result of the interaction of the working fluid with the rock, which will lead to the dissolution of some minerals and the formation of others, with concomitant changes in porosity and permeability. Geochemistry is expected to play an important role in dealing with these issues.

Bibliography

Primary Literature

1. von Knebel W (1906) Studien in den Thermengebieten Islands. Naturwissenschaftliche Rundschau
2. Thorkelsson T (1910) The hot springs of Iceland. Det kongelige danske Videnskabernes Seleskabs Skrifter
3. Ingersoll LR, Zobel OJ (1913) Mathematical theory of heat conduction. Ginn, Waltham
4. Einarsson T (1942) Ueber das Wesen der heissen Quellen Islands. Soc Sci Islandica 42:91
5. White DE (1957) Magmatic, connate and metamorphic waters. Geol Soc Am Bull 69:1659–1682
6. Ellis AJ, Wilson SH (1960) The geochemistry of alkali metal ions in the Wairakei hydrothermal system. N Z J Geol Geophys 3:593–617
7. Ellis AJ, Mahon WAJ (1977) Chemistry and geothermal systems. Academic, New York, p 392
8. White DE (1967) Some principles of geyser activity, mainly from Steamboat Springs, Nevada. Am J Sci 265:641–684
9. Giggenbach WF (1988) Geothermal solute equilibria. Derivation of Na-K-Mg- Ca geoindicators. Geochim Cosmochim Acta 52:2749–2765
10. Giggenbach WF, Le Guern F (1976) The chemistry of magmatic gases from Erta'Ale Ethiopia. Geochim Cosmochim Acta 40:25–30
11. Giggenbach WF (1987) Redox processes governing the chemistry of fumarolic gas discharges from White Island, New Zealand. Appl Geochem 2:143–161
12. Giggenbach WF (1990) Water and gas chemistry of Lake Nyos and its bearing on the eruptive process. J Volcanol Geoth Res 42:337–362
13. Truesdell AH, Fournier RO (1977) Conditions in the deeper parts of the hot spring systems of Yellowstone National Park, Wyoming, US Geological Survey Open-file report No. 76-428. US Geological Survey, Reston
14. Truesdell AH, Nathenson M, Rye RO (1977) The effects of subsurface boiling and dilution on the isotopic compositions of Yellowstone thermal waters. J Geophys Res 82:3694–3704
15. Giggenbach WF (1978) The isotopic composition of waters from the El Tatio geothermal field. Northern Chile. Geochim Cosmochim Acta 42:979–988
16. Nieva D, Verma MP, Santoyo E, Portugal E, Campos A (1997) Geochemical exploration of the Chipilapa geothermal field, El Salvador. Geothermics 26:589–612
17. Mercado (1975) Movement of geothermal fluids and temperature distribution in the Cerro Prieto geothermal field, Baja California, Mexico. In: United Nations Symposium on the Development and Use of Geothermal Resources, San Francisco, 1975, vol 1. U. S. Government Printing Office, Washington, pp 487–494
18. Ellis AJ, Mahon WAJ (1967) Natural hydrothermal systems and experimental hot water/rock interactions (Part II). Geochim Cosmochim Acta 31:519–538
19. Fournier RO, Truesdell AH (1973) An empirical Na-K-Ca geothermometer for natural waters. Geochim Cosmochim Acta 37:1255–1275
20. Truesdell AH (1975) Geochemical techniques in exploration. In: National Energy Authority (ed) Proc. 2nd. UN Symp. on the development and use of geothermal resources, San Francisco, 1975, vol 1. National Energy Authority, Reykjavik, pp 53–86
21. Arnorsson S, Gunnlaugsson E, Svavarsson H (1983) The chemistry of geothermal waters in Iceland III. Chemical geothermometry in geothermal investigations. Geochim Cosmochim Acta 42:567–577
22. Fournier RO (1979) A revised equation for the Na/K geothermometer. Geoth Res Counc Trans 3:221–224
23. Giggenbach WF, Gonfiantini R, Jangi BL, Truesdell AH (1983) Isotopic and chemical composition of Parbati Valley geothermal discharges, NW-Himalaya, India. Geothermics 12:199–222

24. Nieva D, Nieva R (1987) Developments in geothermal energy in Mexico – Part Twelve. A cationic composition geothermometer for prospecting of geothermal resources. Heat Recovery Syst CHP 7:243–258
25. Bowers TS, Jackson KI, Helgeson HC (1984) Equilibrium activity diagrams. Springer, Berlin
26. Fournier RO (1991) Water geothermometers applied to geothermal energy. In: D'Amore F (ed) Application of geochemistry in geothermal reservoir development, Series of technical guides on the use of geothermal energy. UNITAR/UNDP, New York, p 37
27. Paces T (1975) A systematic deviation from Na-K-Ca geothermometer below 75°C and above 10^{-4} atm P_{CO2}. Geochim Cosmochim Acta 39:541–544
28. Fouillac C, Michard G (1977) Sodium, potassium, calcium relationships in hot springs of Massif Central. In: Pacquet H, Tardy Y (eds) Proc. 2nd. Intl. Symp. on Water-Rock Interaction, Strasbourg, 17–25 August 1977, vol 3. Université Louis Pasteur, Strasbourg, pp 109–116
29. Fournier RO, Potter RW (1979) Magnesium correction to the Na-K-Ca chemical geothermometer. Geochim Cosmochim Acta 43:1543–1550
30. Truesdell AH, Nakanishi S (2005) Chemistry of neutral and acid production fluids from the Onikobe geothermal field, Miyagi prefecture, Honshu, Japan. In: IAEA (ed) Use of isotope techniques to trace the origin of acidic fluids in geothermal systems, Technical document 1448. IAEA, Vienna, p 197
31. Baca Gómez A, Segovia N, Martínez Miranda V, Armienta MA, Barragán Reyes RM, Iturbe García JL, López Muñoz BE, Seidel JL (2006) Physical, chemical, bacteriological and radioisotopic parameters from springs and wells around Jocotitlán volcano, Mexico. Int J Environ Pollut 26:266–283
32. Kharaka YK, Mariner RH (1989) Chemical geothermometers and their application to formation waters from sedimentary basins. In: Naeser ND, McCollon TH (eds) Thermal history of sedimentary basins. Springer, New York, pp 99–117
33. Fouillac C, Michard G (1981) Sodium/lithium ratios in water applied to geothermometry of geothermal reservoirs. Geothermics 10:55–70
34. Fournier RO, Rowe JJ (1966) Estimation of underground temperatures from the silica content of water from hot springs and wet-steam wells. Am J Sci 264:685–697
35. Kennedy GC (1950) A portion of the system silica-water. Econ Geol 45:629–653
36. Morey GW, Fournier RO, Rowe JJ (1962) The solubility of quartz in water in the temperature interval 25 °C to 30 °C. Geochim Cosmochim Acta 26:1029–1043
37. Fournier RO, Potter RW (1982) A revised and expanded silica (quartz) geothermometer. Geoth Res Counc Bull 11:3–12
38. Truesdell AH, Fournier RO (1977) Procedure for estimating the temperature of a hot-water component in a mixed water by using a plot of dissolved silica versus enthalpy. USGS J Res 5:49–52
39. Fournier RO (1973) Silica in thermal waters: laboratory and field investigations. In: Ingerson E (ed) Proceedings: International Symposium on Hydrogeochemistry and Biogeochemistry, Tokyo, 7–9 September 1970, vol 1. Clarke, Washington, DC, pp 122–139
40. Rimstidt JD, Barnes HL (1980) The kinetics of silica-water reaction. Geochim Cosmochim Acta 44:1683–1699
41. Giggenbach WF, Goguel RL (1989) Collection and analysis of geothermal and volcanic water and gas samples, Report No. CD 2387. DSIR, Petone, p 53
42. Nieva D, Quijano León JL (1991) Applications of geochemistry to the study of geothermal resources in Mexico: case study of Los Azufres field. In: D'Amore F (ed) Application of geochemistry in geothermal reservoir development, Series of technical guides on the use of geothermal energy. UNITAR/UNDP, New York, pp 299–316
43. Giggenbach WF (1980) Geothermal gas equilibria. Geochim Cosmochim Acta 44:2021–2032
44. D'Amore F, Celati R (1983) Methodology for calculating steam quality in geothermal reservoirs. Geothermics 12:129–140
45. Nieva D, Fausto J, González J, Garibaldi F (1982) Afluencia de vapor a la zona de alimentación de pozos de Cerro Prieto I. In: Fourth Symp. on the Cerro Prieto Field, Baja California, Mexico, vol 1, p 145
46. D'Amore F, Truesdell AH (1985) Calculation of geothermal reservoir temperatures and steam fraction from gas compositions. Geoth Res Counc Trans 9(1):305–310
47. Arnorsson S, Gunnlaugsson E (1985) New gas geothermometers for geothermal exploration – calibration and application. Geochim Cosmochim Acta 49:1307–1325
48. Taran Y (1986) Gas geothermometers for hydrothermal systems. Geochem Int 20:111–126
49. Giggenbach WF (1991) Chemical techniques in geothermal exploration. In: D'Amore F (ed) Application of geochemistry in geothermal reservoir development, Series of technical guides on the use of geothermal energy. UNITAR/UNDP, New York, p 119
50. Cathelineau M, Nieva D (1985) A chlorite solid solution geothermometer. The Los Azufres (Mexico) geothermal system. Contrib Mineral Petrol 91:235–244
51. Cathelineau M (1988) Cation site occupancy in chlorites and illites as a function of temperature. Clay Miner 23:471–485
52. Kranidiotis P, MacLean WH (1987) Systematics of chlorite alteration at the Phelps Dodge massive sulfide deposits, Matagami, Quebec. Econ Geol 82:1898–1911
53. Battaglia S (1999) Applying X-ray geothermometer diffraction to a chlorite. Clays Clay Miner 47:54–63
54. Craig H (1961) Standard for reporting concentrations of deuterium and oxygen-18 in natural waters. Science 133:1833
55. Craig H (1961) Isotopic variations in meteoric waters. Science 133:1702–1703
56. Taylor HP (1974) The application of oxygen and hydrogen isotope studies to problems of hydrothermal alteration and ore deposition. Econ Geol 69:843–883
57. Craig H (1963) The isotopic geochemistry of water and carbon in geothermal areas. In: Tongiorgi E (ed) Nuclear geology on geothermal areas: Spoleto 1963. Consiglio Nazionale della Richerche, Pisa, pp 17–53

58. Giggenbach WF (1989) The chemical and isotopic position of Ohaaki field within the Taupo Volcanic Zone. In: Proc. 11th New Zealand Geothermal Workshop, Aukland, 1989, pp 81–88
59. Giggenbach WF (1991) Isotopic composition of geothermal water and steam discharges. In: D'Amore F (ed) Application of geochemistry in geothermal reservoir development, Series of technical guides on the use of geothermal energy. UNITAR/UNDP, New York, pp 253–273
60. Sakai H, Matsubaya O (1977) Stable isotope studies of Japanese geothermal systems. Geothermics 5:97–124
61. Taran YA, Pokrovsky BG, Esikov AD (1988) Deuterium and oxygen-18 in fumarolic steam and amphiboles from some Kamchatka volcanoes: "Andesitic" waters. IAVCEI, Commission on the Chemistry of Volcanic Gases Newsletter No. 1, pp 15–18
62. Fournier RO (1979) Geochemical and hydrologic considerations and the use of enthalpy-chloride diagrams in the prediction of underground conditions in hot- spring systems. J Volcanol Geoth Res 5:1–16
63. Giggenbach WF, Stewart MK (1982) Processes controlling the isotopic composition of steam and water discharges from steam vents and steam- heated pools in geothermal areas. Geothermics 11:71–80

Books and Reviews

D'Amore F (ed) (1991) Application of geochemistry in geothermal reservoir development. Series of technical guides on the use of geothermal energy. UNITAR/UNDP, New York

Henley RW, Truesdell AH, Barton PB Jr (eds) (1984) Fluid-mineral equilibria in hydrothermal systems. Reviews in Economic Geology, vol I

IAEA (1983) Guidebook on nuclear techniques in hydrology. International Atomic Energy Agency, Technical Report Series No. 91, Vienna

Valley JW, Taylor HP Jr, O'Neil JR (1986) Stable isotopes in high temperature geological processes. Reviews in Mineralogy, vol 16